INDIVIDUAL DIFFERENCES IN SYNTACTIC PROCESSING DURING READING: A PSYCHOLINGUIST'S "TWO DISCIPLINES" PROBLEM

BY

ARIEL N. JAMES

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Psychology
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2017

Urbana, Illinois

Doctoral Committee:

Professor Gary S. Dell, Chair
Associate Professor Duane G. Watson, Vanderbilt University, Director of Research
Associate Professor Sarah Brown-Schmidt, Vanderbilt University
Professor Aaron S. Benjamin
Professor R. Chris Fraley

ABSTRACT

Psycholinguists have identified syntactic structures that are consistently more difficult to read than others. To understand why readers find these structures difficult (and thus, what mechanisms underlie syntactic processing in these contexts), one line of research has sought to link individual differences in reading to individual differences in cognitive abilities. Put another way: how do cognitive differences between readers interact with syntactic processing effects observed across readers? This dissertation describes a single study in which 133 young adults read sentences via a self-paced moving window paradigm and then completed a battery of 16 tasks to assess their abilities in the following areas: language experience, phonological ability, working memory, inhibitory control, and perceptual speed. Three syntactic phenomena were chosen for the current investigation: the relative processing difficulty for object- versus subject-extracted relative clauses; the effect of verb biases in reading a sentential complement; and the tendency to resolve relative clause attachment ambiguities to low attachment sites. Each of these effects is well documented in the psycholinguistic literature, and each has been implicated in processing theories that predict effects of individual differences between adult readers. In both a multi-level mixed-effects regression analysis (1A) and a latent variable analysis (1B), we find correlations between measures of individual differences (notably language experience and memory span scores) and overall reading comprehension, reading speed, and relative clause attachment ambiguity resolution (lower working memory is associated with a high attachment preference). Experimental effects on reading time were not consistent measures within individual subjects, which we suggest limits their ability to correlate with other measures and might explain controversy in the literature over how individual differences are linked to language processing.

For my family, my students, and Ariana

ACKNOWLEDGEMENTS

I would like to highlight some of the people who have supported my work and given me strength and encouragement during the six years I have spent in pursuit of my doctoral degree:

For their supervision of my thesis research, I thank my advisor Duane Watson for supporting my entire graduate career; Gary Dell, who accepted the roles of committee chair, local advisor, and graduation "hooder" in my final year when Duane moved to Vanderbilt; and Sarah Brown-Schmidt, Aaron Benjamin, and R. Chris Fraley who have provided me with helpful feedback as members of my thesis committee. Thank you for your wisdom, and for believing in my work when my own doubts threatened my progress.

For their direct involvement in this work, I thank Scott Fraundorf, Eun-Kyung Lee, and Duane Watson who are co-authors on presentations and manuscripts on this project.

For their service on my qualifying exam committee, I thank Duane Watson, Kara Federmeier, Elizabeth Stine-Morrow, and Steven Culpepper.

For helpful conversations and feedback related to this work, I thank Jennifer Arnold, Gerry Altmann, Brennan Payne, Joseph Toscano, and William F. Brewer; attendees of Aaron Benjamin's lab meeting, the Language Comprehension Lab Meeting (including Cynthia Fisher, Kara Federmeier, Duane Watson, Sarah Brown-Schmidt, and their students), the Cognitive Division Brown Bag and Language Processing Brown Bag; and reviewers at the CUNY Conference for Human Sentence Processing and at JEP: General, including Benjamin Swets and anonymous scholars.

For their assistance in all practical matters, I thank Psychology Department staff members, especially Ashley Ramm, Jim Clark, and Firmino Pinto; lab managers Dominique

Simmons, Loretta Yiu, and Daniela Avelar; and research assistants Sarah Bopp, Bailey Cation, Gabrielle Smith, and Sean Zolfo.

For financial support, I thank the Graduate College Fellowship at Illinois and the Nation Science Foundation Graduate Research Fellowship (DGE-1144245).

For their contributions to my mental health, I thank Dr. Greg Lambeth for his unwavering commitment and life-altering insight; Dr. Marybeth Hallett and members of the women's therapy group; Dr. Pearson; and the Disability Resources & Educational Services center at Illinois.

For providing me with a creative outlet, I thank cellist and cello instructor Sam Araya; the Big Grove Cello ensemble; musicians Venanzio, Carlo, Henry, Caterina, Sergio, and Apoorv; and artist Ron Karlstrom (1950-2014).

For moral support, I thank my family, particularly my parents Bernita, Everett, and Ronda; my brother; my grandparents; and the Anderson, Sogge, Cooke, Foster, James, and Hughes families. I also thank my wonderful friends, old and new, especially the two who were close beside me from my first year in Urbana until my last: my lab mate and academic twin brother Andrés, and Abigail, a rare friend who is both old and new.

Finally, for everything else, I thank Nate Anderson.

TABLE OF CONTENTS

CHAPTER 1: INTRODUCTION

In Lee Cronbach's famous presidential address to the American Psychological Association Annual Convention in 1957, he described an optimistic vision of the future of psychology in which the best of the correlational and experimental traditions joined forces as the *united discipline*. A complete theory of human behavior, he argued, requires the modeling of individual variability along with the prediction of an individual's response to varying conditions. The usefulness of such a united approach is especially clear in the domains of applied psychology: It would be best to provide an intervention that is tuned to the particular needs of each individual (Pellegrino, Baxter, & Glaser, 1999). Since that 1957 address, psychologists have taken up the challenge of the united discipline. In their 1999 review, Pellegrino, Baxter, and Glaser chart the progress of the field, focusing on the intersections of cognitive psychology and psychometrics that follow directly from Cronbach's initial interests, focusing first on "aptitude-treatment interactions", or the relationship between a student's intellectual abilities and expertise on one hand, and educational materials and instructional methods on the other.

A specific case of this type of investigation is what we will call "reader-text interactions". Substantial prior work has revealed that the time required to read a sentence or text is a function of both the individual reader and the text being read: Researchers in individual differences and educational psychology have identified important sources of variation in reading and comprehension skill (e.g., Kuperman & Van Dyke, 2011; Perfetti & Hart, 2002), and work in cognitive psychology and psycholinguistics has identified the types of words, sentence, and texts that are more difficult for comprehenders (e.g., some syntactic structures are more difficult to process; Just & Carpenter, 1992; Waters & Caplan, 1996; Gibson, 1998). What is less clear is whether and how reader and text characteristics *interact*: Are difficult sentences equally

challenging for all readers? And, conversely, does variation in reading skill affect the comprehension of *all* linguistic materials, or just especially difficult ones?

Understanding reader-text interactions furthers our understanding of several broader issues in psychology. First, by understanding how variability in readers interacts with properties of texts, we can gain more general insights into the underlying mechanisms of language processing (for a discussion of how individual differences contribute to more general theoretical development, see Vogel & Awh, 2008.) As we review below, theories of language processing make different claims about why some texts are more difficult to process. Consequently, they also imply different hypotheses about which individual differences are likely to modulate syntactic processing. For instance, theories that attribute the difficulty of some syntactic structures to comprehenders' relative inexperience with them predict that individual differences in language experience might drive differences in syntactic processing. By contrast, in accounts in which some syntactic structures are difficult because of the demands they place on memory, it is individual differences in memory capabilities that are most likely to relate to individual differences in syntactic processing.

Second, individual differences in syntactic processing speak to broader, fundamental questions about the architecture of the mind and language processing system. For example, as we review in greater detail below, some theories (e.g., Waters & Caplan, 2003) propose that language processing is divided into initial, automatic stages and later, interpretive stages, with only the latter subject to individual differences in working memory and other cognitive abilities. Studying individual differences in both online and offline processing allows us to test this theoretical claim. Similarly, another central question in psychology is the extent to which cognitive systems are modular rather than driven by domain-general systems (see Fodor, 1983).

By understanding whether variability in the capacity of domain general systems like working memory and executive function is associated with syntactic processing, we can better understand the overall architecture of the mind: To what degree is language (and other motor and perceptual systems) modular, and to what degree does it recruit domain-general systems? It also provides an opportunity to understand *why* characteristics like high working memory are associated with positive outcomes in more complex domains like reading comprehension.

Finally, and most broadly, reader-text interactions exemplify one of the central questions of the united discipline envisioned by Cronbach: How do the skills and abilities identified by psychometricians intersect with the cognitive-processing effects discovered by experimentalists? Aptitude-treatment interactions have been reported in some educational domains. For instance, learners with greater prior knowledge learn better from different types of texts (McNamara, Kintsch, Songer, & Kintsch, 1996) and feedback (Hausmann, Vuong, Towle, Fraundorf, Murray, & Connelly, 2013) than do low-knowledge learners. Indeed, several reader-text interactions have been reported within the language processing literature. For instance, slower overall readers show larger effects of word frequency (Seidenberg, 1985), and readers with greater linguistic experience may be less sensitive to word difficulty and correspondingly more sensitive to discourse-level factors (e.g., the introduction of new concepts; Stine-Morrow, Soederberg Miller, Gagne, & Hertzog, 2008). Most relevant for the present paper, readers with greater linguistic knowledge are also more efficient at resolving syntactic ambiguity (Traxler & Tooley, 2007). On the other hand, a review of the learning-styles hypothesis—that certain learners do best under one instructional method and other learners do best with a different method—has found little evidence to date in favor of such an interaction; instead, the most well-established mnemonic effects appear to apply across learners (Pashler, McDaniel, Rohrer, & Bjork, 2008). Thus, there

is a need to investigate in other domains whether the cognitive-processing effects discovered by experimentalists are consistent across individuals, and whether the important skills and abilities identified in psychometrics apply across tasks and materials.

Assessing Reader-Text Interactions

Reader-text interactions have been studied by both educational psychologists and cognitive psychologists. While educational psychologists have investigated reader-text interactions with the goal of promoting learning in young readers (e.g. Coté, Goldman, & Saul, 1998) and comprehension among students (e.g., McNamara et al., 1996), a complementary literature grew in cognitive psychology as theories of reading began to include ideas about individual differences in cognitive abilities. An influential example is Just and Carpenter (1992), who proposed, and reviewed evidence, that differences in capacity between individuals correlate with differences in reading ability. Since then, psycholinguists have employed individual differences to promote both memory-capacity theories of language comprehension (e.g. Fedorenko, Gibson, & Rohde, 2006; 2007; Gibson, 1998; 2000), competing experience-based theories (MacDonald & Christiansen, 2002, discussed in greater detail below), and a number of other explanations that combine language-specific and domain-general mechanisms (e.g. Farmer, Fine, Misyak, & Christiansen, 2016; Novick, Trueswell, & Thompson-Schill, 2010; Payne, Grison, Gao, Christianson, Morrow & Stine-Morrow, 2014; Swets, Desmet, Hambrick, & Ferreira, 2007; Van Dyke, Johns, & Kukona, 2014).

As the individual differences approach in psycholinguistics has continued to grow in popularity in recent years, it is important to take a step back and assess its progress toward the *united discipline*. These psycholinguistic investigations are nested within the experimental approach, investigating language-processing effects that have been previously shown across

subjects using controlled linguistic stimuli. So, the question is whether these investigations live up to the ideals of the *correlational* approach. Here, we describe several methodological demands identified by the correlational approach and discuss how these constraints may have contributed to a lack of consensus regarding individual differences in syntactic processing.

First, a critical insight from measurement theory is that two variables can be observed to correlate only to the degree that there is meaningful variation in those individual variables and to the degree that such variation is reliably measured (Spearman, 1904). If there are genuine, stable individual differences in syntactic processing, those individuals who show large syntactic-processing effects on one subset of items should also show large effects on another, similar subset. By contrast, a failure to observe such correlations would suggest that either (a) there are not consistent individual differences in syntactic processing or (b) such differences exist, but our methods cannot reliably detect them. We revisit these alternatives in the Discussion.

For instance, consider a scenario in which all readers read a syntactically complex sentence 300 ms more slowly than a syntactically simple sentence. In this case, there is clearly a *text* effect—one sentence is more difficult than another—but there is no reader-text *interaction* because *all* readers found the complex sentence more difficult than the simple sentence to the same degree. In this scenario, it would be impossible for any other construct (such as verbal working memory) to explain individual differences in syntactic processing because such variation was not observed to begin with. Unfortunately, while past investigations of individual differences in syntactic processing have sometimes used measures of working memory and other cognitive abilities that have been normed for their reliability, researchers have only rarely assessed whether we observe meaningful variation across individuals in the syntactic processing effects themselves (but see Swets, et al., 2007 for one application of psychometric principles to

syntactic processing). Thus, before we ask *why* individuals might differ in syntactic processing, it is first necessary to establish that such individual differences *exist* at all. If we cannot observe consistent individual differences in syntactic processing to begin with, differences in online syntactic processing cannot be expected to relate to any other measure.

Second, individual differences are best assessed with multiple measures. "Perhaps the most valuable trading of goods the correlator can offer," Cronbach (1957) states, "…is his multivariate conception of the world. No experimenter would deny that situations and responses are multifaceted, but rarely are his procedures designed for a systematic multivariate analysis" (p. 676). A strength of the multivariate approach is that it deals explicitly with *measurement error*: Observed performance on almost any single task reflects not only the construct of interest but also measurement error, which includes both random error and non-random error from other constructs (Bollen, 1989). Consider reading span (Daneman & Carpenter, 1980), which has been used as the single measure of verbal working memory capacity in several influential psycholinguistic studies of individual differences (e.g. Just & Carpenter, 1992; MacDonald, Just, & Carpenter, 1992; Pearlmutter & MacDonald, 1995). The reading span task purports to measure verbal working memory capacity because it requires participants to remember particular words while reading sentences, but it might also be influenced by participants' knowledge of specific lexical items (Engle, Nations, & Cantor, 1990; MacDonald & Christiansen, 2002). These confounds make it difficult to interpret a high or low score on any single measure. But, including multiple measures of a single construct allows researchers to assess the degree of common variance between them and use composite scores within a construct; for instance, a composite score for verbal working memory can be created by administering both a reading span and an

operation span task. Unfortunately, not all psycholinguistic studies have used multiple measures (or *indicators*) for any given factor.

Further, the psychometric approach implies that multiple constructs should be measured simultaneously in order to tease apart their effects. A challenge for studying individual differences is that many potential explanatory constructs, such as verbal working memory and linguistic experience, might be intercorrelated (e.g., MacDonald & Christiansen, 2002), making it more challenging to attribute effects to any one construct in particular. In order to demonstrate that a specific construct—say, linguistic experience—is the one that drives differences in online syntactic processing, it is important to also measure other competing constructs and to show that it is specifically linguistic experience, and not (for example) verbal working memory or inhibitory control, that relates to individual differences in processing. However, many psycholinguistic studies have examined only one or two of these constructs within a single investigation; for instance, a study may measure verbal working memory but not reading experience, or vice versa.

The current study aims to address these three issues by (1) assessing multiple constructs—both domain-general and language-specific—within individuals, (2) including multiple measures of each construct (e.g., multiple span tasks to create a composite measure of verbal working memory), and (3) assessing whether our data include consistent individual differences in the predictor variables and in the syntactic processing effects.

CHAPTER 2: DIFFERENCES IN SYNTACTIC PROCESSING (STUDY 1A)

In order to investigate reader-text interactions, this investigation focuses on three different syntactic constructions that have been widely studied in the psycholinguistic literature, investigating potential individual differences both in initial processing and in subsequent comprehension. Five candidate predictors of individual differences—namely, language experience, phonological ability, verbal working memory, inhibitory control, and processing speed—are each measured with multiple tasks in the same sample of subjects, allowing multiple explanations to be tested simultaneously. A review of these explanations is presented below.

What Might Account for Individual Differences in Syntactic Processing?

Language experience. Experience-based accounts propose that individual differences in syntactic processing, even those that are correlated with domain-general abilities, are better explained as differences in exposure to various structures (MacDonald & Christiansen, 2002). This claim about individual differences is consistent with broader theories of language comprehension that posit a strong influence of experience on syntactic processing more generally. For example, constraint-based theories of language comprehension (Altmann & Steedman, 1988; MacDonald, 1994; MacDonald, Pearlmutter, & Seidenberg, 1994; Spivey-Knowlton, Trueswell, & Tanenhaus, 1993) propose that language comprehension is fast and accurate because it incorporates numerous probabilistic constraints, including syntactic ones, that comprehenders have learned through their experience with language. Experience-based theories are supported by demonstrations that syntactic structures are read more quickly when they are more frequent or predictable, as determined from either global statistics or those of particular verbs (Garnsey, Pearlmutter, Myers, & Lotocky, 1997; MacDonald & Christiansen, 2002), and even when memory demands are equated (e.g. Levy, Fedorenko, Breen, & Gibson, 2012).

Experience-based accounts are further supported by evidence that online processing of initially difficult structures can be facilitated on the basis of recent laboratory-provided experience with the structures, including both trial-to-trial changes (Arai, van Gompel, & Scheepers, 2007; Thothathiri & Snedeker, 2008; Tooley, Traxler, & Swaab, 2009; Traxler, 2008) and changes over the course of one or more experimental sessions (Farmer, Fine, Yan, Cheimariou, & Jaeger, 2014; Fine, Qian, Jaeger, & Jacobs, 2010; Fine, Jaeger, Farmer, & Qian, 2013; Wells, Christiansen, Race, Acheson, & MacDonald, 2009), and even for structures that are only marginally grammatical (Luka & Barsalou, 2005; Luka & Choi, 2014) or that were previously unfamiliar to the comprehender (Kaschak, 2006; Kaschak & Glenberg, 2004; Fraundorf & Jaeger, 2016).

The examples thus far discuss differences *between* syntactic structures but *within* individuals. But, the claim that syntactic processing is guided by relative experience with different structures also suggests that processing could be influenced by differences among individuals in their relevant linguistic experience: Some individuals may come into the reading task with substantially more or less of the experience that was experimentally manipulated in some of the experiments described above. Thus, for instance, computational simulations suggest that rare, difficult structures are less disruptive for more experienced readers, who have more experience with these uncommon structures (MacDonald & Christiansen, 2002). This prediction has been supported by recent studies in the spoken language processing domain, which have found that individuals with higher vocabulary or higher literacy show facilitation in online, anticipatory language processing in the visual world (e.g. Borovsky, Elman, & Fernald, 2012; Huettig & Janse, 2016; Mishra, Singh, Pandey, & Huettig, 2012; Rommers, Meyer, & Huettig, 2015). There is still relatively little comparable work in the written modality, but Traxler and

Tooley (2007) found that individuals with greater knowledge were less affected by temporary syntactic ambiguity in their online processing.

Phonological ability. Phonological abilities have long been hypothesized to be a major factor in determining reading ability, particularly in acquisition or among poor readers (e.g. Byrne & Letz, 1983; Read & Ruyler, 1985; Sawyer & Fox, 1991; Wagner, Torgesen, & Rashotte, 1999). Experimental manipulations of phonological interference in text (Baddeley, Eldrige, & Lewis, 1981; Keller, Carpenter, & Just, 2003; McCutchen, France, & Perfetti, 1991; Kennison, 2004, to name a few) also suggest a role of phonology in offline syntactic comprehension even among skilled adult readers.

However, fewer studies have investigated effects of phonology during initial, on-line syntactic processing, and those that have yielded mixed evidence. Acheson and MacDonald (2011) found that sentences with embedded relative clauses were made more difficult by phonological overlap between the head noun of the relative clause and a noun embedded within it (e.g. *baker* and *banker*) and between the relative clause verb and main clause verb (e.g. *sought* and *bought*). This overlap effect was larger for object-extracted relative clauses (ORCs; 1a), which are typically more difficult in general, relative to subject-extracted relative clauses (SRCs; 1b), perhaps because in some theoretical accounts, phonological representations could be used to maintain the non-canonical ordering of agent and patient in the ORC (that is, the sought *baker* precedes the seeking *banker*; MacDonald & Christiansen, 2002).

(1a) The baker that the banker sought bought the house.

(1b) The baker that sought the banker bought the house.

However, Kush, Johns, and Van Dyke (2015) present data that suggest that these effects are the result of encoding interference rather than interference with syntactic integration. Indeed,

some theories (e.g., McElree, Foraker, & Dyer, 2003; Martin & McElree, 2008) propose that maintaining serial order is not necessary for comprehension because previous constituents can be directly accessed in memory. Thus, while Van Dyke et al. (2014) found that reading times were related both to vocabulary and to non-verbal memory for serial order, they found no effects of phonological ability.

Whether variation between individuals in phonological ability plays a role in processing is a point of controversy, but it is possible that individual differences between individuals in phonological ability could also influence syntactic processing ability—especially for structures where it may be important to maintain serial order to arrive at the correct meaning of the sentence.

Verbal working memory capacity. As we introduced briefly above, capacity constraints in verbal working memory have figured prominently in research on reader-text interactions. Some theories have proposed that syntactic structures are difficult to process to the extent that they impose greater demands on memory (Fedorenko et al., 2006, 2007; Gibson, 1998, 2000; Just & Carpenter, 1992; King & Just, 1991). For instance, in both the ORC (1a) and SRC (1b) above, the relative pronoun *that* introduces a dependency in which the relative pronoun must eventually be co-indexed with a syntactic gap in the relative clause. In the ORC, this integration occurs later (at *sought*, the reader must recall it was *the baker* who was sought) and requires a longer-distance memory retrieval than in the SRC, in which the gap occurs immediately after *that*. It has been argued (Gibson, 1998, 2000) that these memory demands explain why ORCs are understood more slowly and less accurately. Thus, differences between individuals in their ability to store and retrieve these dependencies may be associated with how much more difficult they find ORCs.

Other theories suggest a second reason that memory abilities may be important to online language processing. Just and Carpenter (1992) propose that individuals differ in their total capacity to consider multiple sources of information; as a result, individuals with lower memory capacity may also be less able to use additional constraints such as semantic plausibility or referential contexts to help resolve a syntactic ambiguity.

Many studies have evaluated both of these predictions by directly relating syntactic processing to individual differences in measures of verbal working memory. These studies have often used *complex span tasks* in which participants receive sets of items to store and remember while completing a concurrent or interleaved processing task. For instance, participants may read sentences while remembering particular words from the sentences (Daneman & Carpenter, 1980). It has sometimes been reported that readers with lower scores on complex span tasks have greater difficulty with online processing of challenging syntactic structures, such as the object-extracted relative clauses described above (King & Just, 1991). However, Waters and Caplan (1996) point out that low-span readers in these studies performed worse overall and were not differentially more affected by syntactic difficulty. Moreover, studies have revealed inconsistent results as to whether low-span participants are actually more or less influenced by semantic and pragmatic information; some results suggest that low- but not high-span subjects see a benefit in online processing when helpful pragmatic cues are present (King & Just, 1991), and others suggest exactly the reverse (Just & Carpenter, 1992; Long & Prat, 2008; Pearlmutter & MacDonald, 1995; Traxler, Williams, Blozis, & Morris, 2005).

As a result, Caplan and Waters (1999) propose that online, automatic language processing and later interpretive processes tap separate resources and that only later, post-interpretive processes are assessed by complex span tasks and other working memory measures.

For instance, differences in verbal working memory significantly relate to performance on object-extracted relative clauses in offline comprehension accuracy but not in online reading time, even when the measures come from the same participants reading the same sentences (Caplan, DeDe, Waters, Michaud, & Tripodis, 2011; Waters & Caplan, 2005). Indeed, although it is unclear whether such measures correspond to *online* reading, complex span performance correlates with offline syntactic processing, as well as reading comprehension more generally (Daneman & Merikle, 1996). For instance, Swets et al. (2007) found that working memory—even when measured using non-verbal complex span tasks—was significantly associated with how participants would interpret a syntactically ambiguous relative clause in offline comprehension questions (see also Payne et al., 2014).

Inhibitory control. Differences in working memory relate closely to another construct that has been proposed to drive individual differences in language processing: attentional control. Recent work (Novick et al., 2010) has examined syntactic processing as a function of domain-general *inhibitory control*, or the ability to resolve conflict between competing internal representations. Inhibitory control may be necessary for syntactic processing because the interpretation that comprehenders initially favor sometimes turns out to be wholly wrong and needs revision. This possibility is suggested by evidence that an initial misparse, even when later ruled out syntactically (Christianson, Hollingworth, Halliwell, & Ferreira, 2001) or revised by a speaker (Lau & Ferreira, 2005), is not always fully suppressed and may continue to influence readers' eventual, offline interpretations. Indeed, online competition may even arise from syntactic structures that are *never* supported globally but that are coherent in the local syntactic context (Tabor, Galantucci, & Richardson, 2004).

In addition to the demands of revising the syntactic structure of a sentence, inhibitory control may be necessary for resolving competition between similar constituents *online* as the sentence unfolds. For example, the online processing difficulty of object-extracted relative clauses may be amplified by semantic (Gordon, Hendrick, & Johnson, 2001) or phonological (Acheson & MacDonald, 2011) similarity between the referents in the sentence. These findings are consistent with theories, both of language comprehension specifically (Lewis, Vasishth, & Van Dyke, 2006) and of memory more generally (Nairne, 2002), in which the primary determinant of short-term remembering is not a fixed storage capacity but rather the degree of interference between items to be remembered.

Thus, differences in the ability to suppress irrelevant information and resolve competition might lead to differences in the speed and accuracy of comprehension, and such correlations have been observed (Novick, Trueswell, & Thompson-Schill, 2005). More generally, the ability to suppress incorrect or irrelevant information has been argued to contribute to many aspects of language comprehension ability (Gernsbacher, 1993). Differences in inhibitory control might even account for effects previously attributed to working memory capacity: Measures of inhibitory control often correlate with complex span task performance, and individual differences in performance on such tasks have sometimes been attributed in whole (Engle, 2002) or in part (Unsworth & Engle, 2007) to differences in inhibitory control. Indeed, it has been proposed that working memory span performance correlates with language comprehension and other complex activities because each of these activities rely on general attentional control processes (for review, see Kane, Conway, Hambrick, & Engle, 2007).

Perceptual speed. The final construct explored here is *perceptual speed*, or how quickly one is able to process perceptual stimuli (in the visual domain, within the current study), an

ability that falls under the more general construct of *processing* speed (Salthouse, 1996). The inclusion of this basic ability is intended to capture and control for shared aspects of the reading task and other cognitive tasks that result from rapid visual processing of on-screen stimuli. For instance, perceptual speed has been proposed as one of the core abilities that support working memory (see Jarrold & Towse, 2006, for review), so controlling for perceptual speed would allow us to examine other aspects of working memory that may relate more to sentence processing. In addition, perceptual speed itself has been implicated in individual differences in language processing, although most frequently as an explanation for age-related changes in cognition (e.g., Salthouse, 1996; Caplan et al., 2011). Nevertheless, individual differences in processing speed may also explain some of the variability *within* an age group.

Current study

In the current study, we examined the contributions of both domain-specific and domain-general mechanisms to online and offline syntactic processing, providing evidence for how multiple facets of a reader's ability interact with comprehension. We assessed individual differences in all five of the above constructs (language experience, phonological ability, verbal working memory, inhibitory control, and perceptual speed) within the same set of participants, allowing for their effects to be distinguished and compared. Further, each of these constructs was assessed with multiple tasks, which allows us to create composite measures and mitigate task-specific effects.

We then examined the influences of these five predictor constructs on syntactic processing. We selected three syntactic constructions that have been relevant in the psycholinguistic literature in motivating both general theories of language processing and specifically those of individual differences. Our choice of constructions also allowed us to

measure both online processing and offline comprehension, which provide insight into potential differences between interpretative and post-interpretive mechanisms. Critically, we also measured the internal consistency of each of these measures: Do we, in fact, observe consistent individual differences such that (for instance) some subjects consistently find ORCs easier to read than do other subjects?

Finally, we applied linear mixed-effects regression to relate the individual differences to syntactic processing. One potential challenge in distinguishing the influences of, say, verbal working memory and language experience is that, with a relatively large number of predictors and too few observations, regression models tend to capitalize on chance aspects of the data rather than yield generalizable results (the problem of *overfitting*; Babyak, 2004). Linear mixed-effects models solve this problem because the unit of analysis is the reading time on an individual word or the response to an individual comprehension question, rather than an average of all of a participant's reading times or responses. Thus, thousands of observations are available to the regression model. (For further discussion of linear mixed-effect models and other solutions to the study of individual differences in reading, see Matsuki, Kuperman, & Van Dyke, 2016).

Below, these three syntactic structures and their corresponding processing measures are described in detail.

Structures of interest

Relative clause extraction. First, we tested differences in reading and comprehending object-extracted versus subject-extracted relative clauses, a hallmark syntactic phenomenon that has contributed to numerous theories of syntactic processing. As reviewed above, within a participant, ORCs are typically more demanding and are read more slowly than SRCs within the

relative clause; to preview, we replicate this well-established effect in our own data. Our interest, however, was whether there were differences *across* participants in the degree to which ORCs were relatively more difficult than SRCs. Thus, we took as a measure of individual differences the degree to which each participant read the syntactically difficult ORCs more slowly than the syntactically simpler SRCs.

Verb bias. We also examined a second widely-studied phenomenon in syntactic processing: the online use of verb distributional statistics in the sentential complement structure. In sentence (2), a temporary ambiguity between a direct object and sentential complement reading is introduced. In (2a), the ambiguity is resolved early: The complementizer *that* signals that the main verb *accepted* takes a sentential complement in which *the contract* is the subject. In (2b), removing the complementizer makes *the contract* temporarily ambiguous between the subject of the sentential complement (the player accepted some fact about the contract) and the direct object of *accepted* (the contract is what the player accepted).

(2a) The basketball player accepted that the contract required him to play every game.

(2b) The basketball player accepted the contract required him to play every game.

In general, the verb *accepted* is more likely to take a direct object than a sentential complement. Correspondingly, in the ambiguous version, readers slow down when the sentence is disambiguated to the sentential complement structure (at the verb *required*), suggesting they had initially favored the direct object interpretation that is consistent with the distributional statistics of *accepted*. However, other verbs, such as *acknowledged*, take a sentential complement more than a direct object; for these verbs, there is no benefit to disambiguating the structure with *that*, suggesting that readers already favor the sentential complement interpretation (Fine et al., 2010;

Garnsey et al., 1997; Wilson & Garnsey, 2009; but see Kennison, 2001). Thus, our dependent measure of interest was individual differences in magnitude of this verb bias x ambiguity interaction, which indexes the influence of these distributional statistics on online syntactic processing. The use of verb bias is of interest not only because it is another cue that is available during online processing, but because the learning of these biases provides evidence for how processing is shaped through experience with the language environment (for further discussion, see Ryskin, Qi, Duff, & Brown-Schmidt, 2016).

Attachment ambiguity. Finally, we examined the resolution of globally ambiguous relative clause attachments, such as (3) below:

> (3) The maid of the princess who scratched herself in public was terribly embarrassed.

The relative clause *who scratched herself in public* could modify either *the maid* or *the princess*. No syntactic information within the sentence resolves this ambiguity, but attaching the relative clause to the second noun (*low attachment*) is more common than attaching to the first noun (*high attachment*) in English (Rayner, Carlson, & Frazier, 1983), though not in all languages (Cuetos & Mitchell, 1988).

For these items, our interest was purely in participants' offline syntactic processing (in contrast to Payne et al., 2014). Specifically, we queried whether participants arrived at the low attachment or high attachment reading, as revealed by offline probe questions, such as *Did the princess scratch herself*? Note that a "yes" answer to this question, taken alone, might reflect either a genuine low-attachment preference or a simple bias to affirm whichever interpretation is presented. However, as detailed in the Method and Results sections, we varied the question type across items, which allowed us to obtain a measure of participants' low-attachment preference

that was independent of a bias to respond "yes"; this measure of low-attachment preference then served as the key individual-difference variable for these items.

Research Questions

For each of these structures, we considered three questions. Our first question was simply whether we in fact observe consistent individual variation in each of the syntactic processing effects described above. That is, are there some individuals who are consistently advantaged at reading ORCs relative to other individuals? Do some individuals consistently show a stronger low-attachment preference than others? As we note above, a critical first step is to establish that individual differences *exist* and have been reliably measured before considering what other constructs might explain those differences. However, although many studies have sought to relate verbal working memory and other such constructs to online sentence processing, researchers have not always assessed whether there are genuine individual differences in sentence processing to begin with.

Where we found that individuals do vary significantly in their syntactic processing, our second question was determining which individual differences, if any, relate to this variability: Are they domain-specific influences such as linguistic experience, or are they more domain-general abilities such as verbal working memory or executive function?

Finally, we considered whether the relationship between sentence processing and any of the individual differences here is present only in online processing, only in offline comprehension, or in both. Caplan and Waters (1999) propose that there are different constraints on online versus post-interpretive processing, and that only the latter are sensitive to differences in capacity between individuals; however, direct tests of this claim have still been relatively sparse in the literature.

Method

Participants

One hundred and thirty-three subjects participated for course credit or a cash honorarium. The study was advertised to the campus community and was thus biased toward younger adults and university students. Of the 133 participants, 10 did not provide any demographic information: Nine did not show up for the second session, in which the questionnaire was given, and one declined to complete the questionnaire. Of the 123 participants with demographic information, 78 (63%) were female. Participants' ages ranged from 18 to 67 years (M = 20.94 years; SD = 5.37; median = 20 years; 94.3% under age 30). Our sample had only slightly more years of formal education than the nationwide mean (M =13.3 years completed; SD = 1.91; median = 12 years; range = [12, 19]; versus a nationwide mean of 12.9 years according to the United Nations Development Programme, 2014). Most participants (87%) indicated that they had completed at least "some college," and of the 16 remaining responses, 10 came from University students participating for course credit, who presumably did in fact have some college education.

All participants reported that they were native speakers of English who had not been exposed to any other languages before the age of 5 and that they had normal or corrected-to-normal vision and hearing.

Materials

Critical stimuli for the self-paced moving window task consisted of 80 sentences with DO- or SC- bias verbs, 32 unambiguous subject-modifying relative clause sentences manipulated for extraction type, and 20 globally ambiguous relative clause sentences. We describe each of these stimulus types in detail below.

Use of verb bias. The online use of verb bias was tested using 80 critical sentences taken from Lee, Lu, and Garnsey (2013). Each sentence included a matrix subject, followed either by a DO-bias verb (40 sentences) or by a SC-bias verb (40 sentences), and then followed by a sentential complement. Each sentence had 2 versions that differed from each other solely in whether the sentential complement was headed by the complementizer *that*. Example sentences are presented in (4) below. (Emphasis is added here for illustration purposes only and was not presented to participants.).

(4a) DO-biased verb: The club members **understood** (that) the bylaws would be applied to everyone.

(4b) SC-biased verb: The ticket agent **admitted** (that) the mistake might be hard to correct.

In the version without *that*, the role of the post-verbal noun was temporarily ambiguous between the direct object of the verb and the subject of a sentential complement. This ambiguity persisted until the next word (e.g., *would* in 4a or *might* in 4b), which disambiguated the sentence towards a sentential complement structure. In the version with the complementizer, the post-verbal noun was unambiguously the subject of a sentential complement.

Lee et al. (2013) controlled the character length and Francis-Kucera log word frequency of the post-verbal noun across verb type. Although the post-verbal noun was intended in all cases to be highly plausible as a direct object of the verb, plausibility as a direct object was rated as slightly higher after DO-bias verbs than after SC-bias verbs in a norming study conducted on a 7-point scale (6.4 and 6.1 respectively; 1: highly implausible, 7: highly plausible). For details of these norms, see Lee et al. (2013).

We used self-paced reading times to measure participants' online processing of the verb bias items. For both sentence types, the critical region of analysis consisted of the embedded verb and the word immediately afterward, such as *would be* or *might be*, underlined in 4a and 4b above ("the disambiguation region" following Garnsey et al., 1997).

To measure offline comprehension, we created a YES-NO comprehension question for each sentence measuring participants' understanding of its general meaning (e.g., *Did the ticket agent think the mistake would be a problem?*). The questions did not probe whether the participant arrived at the direct object or sentential complement interpretation.

Subject- versus object-extracted relative clauses. Processing of subject- versus object-extracted relative clauses was examined using 32 critical items taken from Gibson, Desmet, Grodner, Watson, and Ko (2005). Critical items began with a subject noun phrase, which was modified by a relative clause, and then continued with the verb phrase of the main clause of the sentence. Each item was manipulated for relative clause extraction site as shown in (5) below. The antecedent noun (in this case, *reporter*) was the subject of the relative clause in the SRC condition, and it was the object of the relative clause in the ORC condition.

(5a) SRC: The reporter <u>who attacked the senator</u> on Tuesday ignored the president.

(5b) ORC: The reporter <u>who the senator attacked</u> on Tuesday ignored the president.

Because the order of the words in the relative clause differed across extraction type, self-paced reading times were analyzed for a combined region including all of the relevant words (following Gibson et al., 2005). This region is underlined above and consisted of the relative pronoun *who*, the noun phrase, and the verb.

For each item, a YES-NO comprehension question was also created to assess offline comprehension. In half of the items, the questions required identifying the subject and object of the relative clause correctly (*e.g., Did the reporter attack the senator? / Did the senator attack the reporter?*). In the other half, the questions asked about main clauses (*e.g., Did the reporter ignore the president? / Did the senator ignore the president?*). This distinction allowed us to probe whether any difficulties in interpreting the ORCs were driven by difficulty in interpreting the relative clause in particular as opposed to the sentence more broadly.

Offline resolution of relative clause attachment ambiguities. To test offline judgments of relative clause attachment, we used 20 relative clause sentences taken from Swets et al. (2007). Each sentence contained a complex noun phrase modified by a relative clause, which was followed by the verb phrase of the main clause. The complex noun phrase included two animate nouns that were linked by the preposition *of*. Relative clauses contained a reflexive pronoun that could refer to either noun of the complex noun phrase, thus creating an attachment ambiguity. An example sentence is presented in (3), reproduced below.

(3) The maid of the princess who scratched herself in public was terribly embarrassed.

For each item, we created a YES-NO question asking explicitly about relative clause attachment. In half of the items, a YES response indicated a low attachment interpretation (e.g., *Did the princess scratch herself?*); in the other half, a YES response indicated a high attachment interpretation (e.g., *Did the maid scratch herself?*). This design allowed us to apply signal-detection analyses (Green & Swets, 1996; Macmillan & Creelman, 2004; Murayama, Sakaki, Yan, & Smith, 2014) to separate participants' potential response bias (any overall tendency to answer *yes* to all questions) from their low-attachment preference (an increase in *yes* responses

specifically when that response indicates a low-attachment reading, termed *sensitivity* in the signal-detection framework).

List construction. Two lists were constructed by counterbalancing the complementizer presence-absence pairings for each of the 80 verb bias sentences and the SRC-ORC pairings for each of the 32 unambiguous relative clause sentences across lists. The stimuli for the 20 globally ambiguous relative sentences were identical across lists. In addition to these 132 experimental sentences, each list contained 80 filler sentences of various structures. The filler sentences were constructed to include a variety of grammatical structures and thereby disguise the structures of interest. Seventeen fillers were passive sentences (e.g., *The terrifying monster was killed by the heroic knight*), twenty-three were simple transitive sentences (*The motivational speaker fixed the projector before her lecture*), six included infinitive clauses (*The game show contestant expected to win*), four were simple intransitive sentences (*The four kids shrieked when the monster appeared on screen*), three were ditransitive sentences (*The friendly man lent sugar to the neighbor next door*), eight were conjoined sentences (*Tania was accepted to graduate school and Steve passed the bar exam*), sixteen used the sentential-complement structure but with a post-verbal noun phrase that was implausible as a direct object of the verb (eleven with the complementizer *that* and five without; e.g., *The housewife hoped the antiques were valuable*), two used the past progressive (*The experienced flight attendant was giving instructions to a group of trainees*), and one was an existential (*There is an old house on the street whose roof was fixed*).

Because the filler sentences were not constructed to be syntactically difficult or confusing, the comprehension questions did not specifically probe the syntax of the sentences but

rather their general semantic content (e.g., *Did Steve fail the bar exam*?). For half of the fillers, the correct answer to the comprehension question was *true*; for the other half, it was *false*.

All participants saw the experimental and distracter sentences in the same, pseudo-randomized order. This design was motivated by our goal of measuring differences between individuals in their language processing, which requires minimizing extraneous sources of variability between participants. Differences in the experimental procedure (e.g., item ordering) across participants introduce additional, irrelevant between-participants variance that cannot be explained by the constructs of interest. By contrast, presenting items in the same order to all participants, although it confounds variance in item properties with serial position, crucially reduces the variance between participants in their experience in the experiment, and the goal of the present study was to explain variance between individuals rather than between items. (See Swets et al., 2007, for another example of an application of this principle to language processing studies.)

Procedure

Participants completed a total of 16 tasks (described individually in detail below) over two experimental sessions 24 hours apart. All participants completed the tasks in the same order to minimize experimental variability between individuals. First, participants completed a self-paced moving-window reading task designed to measure syntactic processing. Participants then completed a battery of tasks measuring the other individual differences of interest. On the first day, these tasks included, in order, three measures of verbal working memory (Reading Span, Listening Span, and Operation Span), two measures of perceptual speed (Letter Comparison and Pattern Comparison), three measures of inhibitory control (Antisaccade, Stroop, and Flanker), and two of five measures of language experience (vocabulary and Author Recognition Test). On

the second day, participants completed a third language experience task (North American Adult Reading Test), three measures of phonological ability (Pseudoword Repetition, Phoneme Reversal, and Blending Nonwords), and finally the two remaining language experience measures (Comparative Reading Habits and Reading Time Estimates questionnaires). Between tasks, the list of tasks was displayed on the screen with checkmarks beside the completed tasks to indicate subjects' progress. Participants were encouraged to take breaks between tasks as needed.

All tasks were completed on a Macintosh desktop computer running MATLAB and the Psychophysics Toolbox (Brainard, 1997; Kleiner, Brainard, & Pelli, 2007; Pelli, 1997). Participants sat approximately 750 mm from the screen.

Self-paced moving window. Syntactic processing was assessed through a self-paced moving-window reading task (Just, Carpenter, & Woolley, 1982). The first word of a sentence was displayed on the screen, with each remaining word in the sentence replaced by a number of dashes equal to the character length of the word (e.g., *chair* would be replaced with -----). When the participant pressed the space bar, the next word was displayed and the previous word was replaced by dashes. Sentences were aligned with the left edge of the screen and displayed equidistant from the top and bottom of the display. All sentences occupied only a single line of text on the screen.

After participants read the last word of a sentence, the sentence disappeared, and a comprehension question was presented in its entirety. Participants answered *yes* or *no* by pressing one of two keys on the keyboard.

Between trials, the serial position of the upcoming trial was displayed for 750 ms in the same screen position as the first word of each sentence. Participants were given a rest period every 40 trials. This task lasted approximately forty-five minutes.

Reading Span. As in all variants of the Reading Span task (Daneman & Carpenter, 1980), participants read sentences while remembering material for a memory test. In the reading portion of the task, participants saw a sentence defining a common noun either truthfully, as in (6a), or falsely, as in (6b). Sentences were taken from Stine and Hindman (1994). Approximately half of the sentences were true and half were false.

(6a) An article of clothing that is worn on the foot is a sock.

(6b) A part of the body that is attached to the shoulder is the toe.

Each sentence was displayed in its entirety in the center of the screen. Participants read the sentence aloud, and then pressed the space bar. The sentence disappeared and was replaced with the prompt "Is this true?" Participants pressed one of two keys on the keyboard to judge the sentence as *true* or *false*.

One goal was to obtain measures of complex span performance that were less influenced by participants' linguistic experience, which otherwise might explain any potential relation between verbal working memory and sentence processing (Engle et al., 1990; Macdonald & Christiansen, 2002). For instance, one way that language experience could influence span scores is by speeding the processing (sentence-reading) component of the task: If all participants saw the sentences for the same amount of time, those participants who could read the sentences more quickly would have more time remaining to implement rehearsal strategies (Friedman & Miyake, 2004). Indeed, allowing participants time to implement strategies in this way reduces the predictive power of complex span tasks (Friedman & Miyake, 2004; McCabe, 2010; Unsworth, Redick, Heitz, Broadway, & Engle, 2009). Thus, we followed the procedure of Unsworth, Heitz, Schrock, and Engle (2005) to reduce the influence of language processing speed by introducing an initial calibration phase to the task. During the initial calibration phase, participants

performed only the processing (semantic judgment) task on 15 sentences and did not perform the memory storage task described below. Participants had unlimited time to read each sentence and make the judgment, and they received feedback on their accuracy afterwards. This procedure was designed to assess each participant's reading speed. We then controlled for reading speed in the main task by giving participants a response deadline that was based on their speed in the calibration phase. In the Results section, we provide evidence that these procedures successfully deconfounded Reading Span scores from language experience.

A second way that language experience might influence complex-span performance is by facilitating processing of the to-be-remembered items. In some versions of the Reading Span task (such as the original version by Daneman & Carpenter, 1980), the to-be-remembered items are the final words of the sentences in the processing task. However, participants' ability to remember such words is influenced by their familiarity or experience with the lexical items themselves (Engle et al., 1990). We thus instead adopted the procedure of Unsworth and colleagues by asking participants to remember letters, which all participants should find highly familiar and easy to process. The letters were randomly chosen from the set *F, H, J, K, L, N, P, Q, R, S, T, Y*, with the constraint that no letter ever appeared twice within the same trial. After each sentence in the main task, the to-be-remembered letter was displayed in caps in the center of the screen for 800 ms.

We also took two other steps to reduce participants' ability to implement strategies. First, participants were required to read the sentence aloud and to press the space bar immediately after doing so; the program displayed a warning if participants were too slow at reading the sentences. Past work has established that stricter pacing of complex span tasks increases their predictive power (Friedman & Miyake, 2004; McCabe, 2010; Unsworth et al., 2009). Second, to prevent

participants from neglecting the reading task in favor of rehearsing the to-be-remembered items, participants were instructed that their primary goal was to maintain at least 85% accuracy on the reading portion of the task. After each test phase, participants saw their cumulative accuracy on the processing task (i.e., their accuracy in judging the sentences as true or false) and received a warning whenever it dropped below 85% (Unsworth et al., 2005).

After completing the calibration procedure, participants proceeded to the main task. Participants continued to read sentences and judge them as true or false, but the maximum time allowed to read a sentence and make the semantic judgment was now set as the participant's mean reading time in the calibration phase plus 2.5 standard deviations (Unsworth et al., 2005). If participants took longer than this time, "TOO SLOW!" displayed on the screen for 1000 ms, the sentence was counted as an error, and the computer proceeded to the next sentence. Participants did not receive feedback on their processing accuracy during the main task. After a predetermined number of sentences and letters, participants proceeded to the test phase of each trial, in which they were required to type the to-be-remembered letters in the order in which they had been presented.

Within the main task, participants first completed two practice trials at span length two (that is, two sentences and a total of two to-be-remembered letters). The critical trials consisted of two trials each at span lengths two to six, for a total of ten trials. A common procedure for complex span tasks has been for participants to start at the shortest span length and progress towards the longest span length, with the task ending if participants do not meet some criterion level of performance. However, researchers have raised several concerns with this procedure. First, performance typically decreases over repeated memory tests (the phenomenon of *proactive interference*). Presenting spans in ascending order confounds span length with the amount of

proactive interference, and so variability in complex span performance could actually reflect variability in susceptibility to proactive interference (Lustig, May, & Hasher, 2001, but see Salthouse & Pink, 2008). Second, concluding the task early reduces the data collected from each participant. Participants may succeed or fail at a particular span lengths for reasons other than their putative verbal working memory abilities, such as the idiosyncratic difficulty of particular sentences (Conway, Kane, Bunting, Hambrick, Wilhelm, & Engle, 2005). Thus, even if a participant does not completely succeed at a given span length, performance at longer spans can still be revealing of their verbal working memory ability. Consequently, we presented the spans in a random order and required all participants to complete all spans.

Scoring was performed according to the *partial-credit unit scoring* procedure recommended by Conway and colleagues (2005). Trials on which participants remembered all of the items were scored as 1 point. Trials on which participants remembered some but not all items were scored as the proportion of items the participants *did* remember. This procedure makes use of all of the information available about participants' performance and incorporates the fact that, for instance, remembering five out of six items represents somewhat better performance than remembering one out of six items. In a comparison of several scoring systems, Conway and colleagues found this procedure to produce the most normal distribution of scores.

Operation Span. The Operation Span task (Turner & Engle, 1989; Unsworth et al., 2005) was also intended to measure verbal memory and generally followed a similar procedure to the reading span task, except that the processing component of the task involved verifying the solutions to equations such as (7).

(7) $(6 \times 4) - 2 = ?$

In the processing portion of the Operation Span task, participants silently read the equation and pressed the space bar when finished. The equation was erased and a probe (such as *22*) displayed on the screen; participants pressed one of two keys to judge whether or not the probe was the correct answer to the equation. Equations were generated according to the procedure of Unsworth et al. (2005). Specifically, the three numbers were always digits between 1 and 9. The first two digits were multiplied or divided together; then, a third digit was added or subtracted. These digits were selected semi-randomly such that the final answer was always a positive integer. Approximately half the test probes were true, and half were false. False probes were generated from the true answer by adding or subtracting a random number between one and nine, with the constraint that the resulting probe was always a positive integer.

As in the Reading Span procedure described above, participants first completed 15 equations in a calibration phase, which involved only the processing component of the task, in order to set the response deadline of the main task. The to-be-remembered items in the main task were the same set of letters used in the reading span task. Participants completed one practice trial at span length two and one at span length three, followed by three critical trials each at span lengths three to seven (for a total of 15 critical trials). As in the Reading Span task, the critical trials were presented in random order.

Listening Span. The Listening Span task generally followed the same procedure as the Reading Span task. However, rather than reading printed sentences aloud, participants listened to pre-recorded sentences spoken by a female native speaker of American English. The prompt to judge the sentence as true or false appeared immediately after the recorded sentence ended. Because the recorded sentence had an identical duration for all participants, calibration of the response deadline was based only on the latency to respond to the prompt. The to-be-

remembered letters were also spoken aloud by the same recorded speaker. The task followed the same procedure as the Reading Span task in all other aspects.

Stimulus sentences were also taken from Stine and Hindman (1994) but comprised a different set of sentences than used in the Reading Span task. There were two practice trials at span length two, followed by two critical trials each at span lengths two to six, again presented in random order.

Letter Comparison. The Letter Comparison task followed Salthouse and Babcock (1991). Participants judged, as quickly as possible, whether two arrays of consonant letters were identical. Trials were presented in six blocks: two blocks comparing three-letter arrays, two blocks comparing six-letter arrays, and two blocks comparing nine-letter arrays. For practice, participants first completed two trials with three-letter arrays, in which one trial contained a match and the other contained a mismatch. Then, during each block, participants were given 20 seconds to complete as many comparison trials as possible, pressing one key for matching arrays and another for mismatching arrays. On mismatching trials, only one letter differed between the arrays. The dependent measure was the total number of correct answers provided within the duration of the critical blocks.

Pattern Comparison. The procedure of the Pattern Comparison task was the same as Letter Comparison, except that participants compared arrays of line segments rather than letters (Salthouse & Babcock, 1991). Blocks of three-, six-, and nine-segment arrays were presented in an order identical to that in the letter comparison task, with the dependent measure being the number of correct answers provided within this time.

Vocabulary. One word was displayed at the top of the screen in capital letters, followed by five other words (in lower case) and *DON'T KNOW*. Participants pressed one of the keys 1-5

on the keyboard to indicate which word was closest in meaning to the word at the top, or they pressed 6 if they did not know. There was one practice item, followed by two critical blocks of 24 items each. Participants had six minutes to complete each block; all participants completed the task within this time limit. All items were taken from the Extended Range Vocabulary Test of the Kit of Factor-Referenced Cognitive Tests (Ekstrom, French, Harman, & Dermen, 1976). Following the procedure recommended by Ekstrom et al. (1976), the dependent measure was the number of correct responses minus a penalty of 0.25 for each incorrect guess. Responses of *DON'T KNOW* were not penalized.

Author Recognition Test. The Author Recognition Test (ART) was developed as a measure of exposure to print materials (Stanovich & West, 1989). We used an updated and slightly lengthened version of the task developed by Acheson, Wells, and MacDonald (2008), which included the names of 65 authors' names and 65 foil names, and adapted that version of the task for the computer. Participants saw names presented one at a time in a random order. For each name, the participant clicked one of two response buttons that appeared at the bottom of the screen reading *Author* and *Don't know*. Participants were told that there was a penalty for guessing, so they were encouraged to only respond with *Author* if they were sure, and to otherwise choose *Don't know*. Participants received one point for each correctly identified author, they lost one point for each foil name that they identified as an author, and there was no change to the score if they selected *Don't know*.

North American Adult Reading Test. The North American Adult Reading Test (NAART) was developed as a way to estimate pre-morbid IQ in brain trauma patients (Blair & Spreen, 1989). Participants received a list of 61 words with irregular spellings, presented one at a time at increasing difficulty. The participants' task was to correctly pronounce each word aloud.

Correct pronunciations, determined by Merriam-Webster's online dictionary, were given one point. Any incorrect response was given zero points with no partial credit. Table 1 displays inter-rater reliability for the NAART and for the other tasks discussed below that require manual scoring.

Comparative Reading Habits (CRH) survey. Participants answered five questions comparing their own self-reported reading habits to what they perceive to be the norm for their fellow college students (Acheson, Wells, & McDonald, 2008).

Reading Time Estimate (RTE) survey. Participants estimated how many hours in a typical week they read various types of materials, including fiction, newspapers, and online materials (Acheson et al., 2008).

Stroop. Following Stroop (1935) and Brown-Schmidt (2009), the Stroop task consisted of two phases. In the first, no-conflict phase, participants named the color of squares displayed one at a time on the screen. The possible colors were red, blue, green, yellow, purple, and orange. Before beginning the task, participants viewed a screen that displayed all six of the possible colors and their names. During the task, participants spoke aloud the name of the color of the square and then pressed a key to advance to the next trial; the key press was used to record participants' response time for the trial.

In the second, conflict phase, participants performed the same task, except that the colored squares were replaced by the English names of colors (e.g. *red* printed in blue). Again, participants' task was to name the color that the word appeared in, rather than read the word aloud.

Each phase contained 100 trials. There was no practice block in either phase, but the first and last trials in each phase were excluded from analysis to account for extreme reaction times attributed to beginning and ending the task.

Participants' responses were recorded and coded for accuracy. Trials were coded as errors if the participant produced the incorrect color name, did not name a color at all, produced a filled pause such as *uh* or *um* (Fraundorf & Watson, 2013; Maclay & Osgood, 1959), or began speaking an incorrect color name before correcting themselves (e.g. *gree- blue*). Accuracy was generally high even in the conflict phase (M = 94%), and all participants obtained accuracy of 74% or greater.

The dependent measure was the difference in median response time between the conflict (second) phase and the no-conflict (first) phase. Because response times were positively skewed, as is typical in response time tasks (e.g., Van Zandt, 2000), response times were first log-transformed before conflict scores were calculated. Only correct trials were analyzed.

Antisaccade. Following the procedures of Kane, Bleckley, Conway, and Engle (2001), participants needed to look in the opposite direction of an anti-predictive cue in order to identify a letter briefly flashed on the opposite side of the screen. Each trial began with a fixation cross that lasted 200, 600, 1000, 1400, 1800, or 2200 ms; this duration varied across trials in order to prevent participants from anticipating the onset of the target. A cue (the equality sign =) then flashed one line of text below the fixation point, at either 11.3 degrees of visual angle to the left or to the right. The cue was visible for 100 ms, disappeared for 50 ms, and reappeared for 100 ms. The target display was then presented at the opposite location (e.g., if the cue appeared on the left, the target appeared on the right) on the same line of text as the fixation point. The target display consisted of a forward mask (the letter H) for 50 ms, then the target letter itself (B, P, or

R) for 100 ms, and then a backward mask (the numeral 8). The backward mask remained on the screen until participants indicated the identity of the target by pressing the 1, 2, or 3 key on the keyboard. All of the characters subtended 2 degrees of visual angle vertically on the screen. There was a 400 ms interval between trials.

Participants first completed 18 trials in a response-mapping phase to practice the mapping between letters and response keys. In this phase, no cue appeared, and the masks and target appeared in the center of the screen. The response mapping was followed by 52 practice trials of the full task. During the practice trials only, participants received feedback in the form of a 175 Hz tone for 500 ms in response to incorrect responses. There was no feedback for a correct response. The practice trials were followed by 72 critical trials. Each possible combination of target identity (B, P, R), target location, and fixation duration was represented twice, and the trials were presented in random order. The dependent measure was the proportion of trials in the critical block on which participants responded correctly.

Flanker. Participants completed a version of the "flankers" response competition paradigm (Eriksen & Eriksen, 1974; see Eriksen, 1995 for review) in which a visually-presented target item is flanked either by congruent items that facilitate correct responding or by incongruent items that inhibit correct responding. In this particular implementation, participants indicated the direction of an arrow that was flanked by four arrows of the same (< < < < <) or different (> > < > >) direction. The incongruent items are thought to activate the incorrect response, making selecting the correct response more difficult, as reflected in longer response latencies (Eriksen, 1995). Similar to the Stroop analysis, the dependent measure was the difference between the median of log-transformed reaction times in the incongruent versus congruent trials.

Pseudoword Repetition. Following Gupta (2003), participants listened to recordings of pseudowords that were phonotactically legal in English (e.g., *ginstabular*), spoken by a female native speaker of American English. After each recording ended, a green dot appeared on the center of the screen and participants attempted to repeat the pseudoword they had just heard. When participants had finished repeating the word, they pressed a key and, after a 100 ms delay, the next trial began. To ensure that participants attempted to produce each word, participants could not end the trial before at least 1000 ms had elapsed; this time point was signaled by the dot on the screen turning blue. There were four critical blocks, each with 18 words: six two-syllable words, six four-syllable words, and six seven-syllable words. Before the main task, participants also completed six practice trials, two at each syllable length. Materials were taken from Gupta (2003).

Participants were awarded one point for each correctly repeated syllable from the onset of the word; correctly repeated syllables that occurred after an erroneous syllable did not earn points. For example, repeating *ginstabular* as *ginstabcular* would score two points; the first two syllables were repeated correctly, but the fourth syllable, while correct, occurred after an error in the third syllable. Some trials (7%) could not be coded because of problems in the recordings, usually because the participant pressed the key before completing the word; for this reason, the dependent measure used was the proportion of points earned out of the points possible on the coded trials only.

Blending Nonwords. Blending Nonwords is a task from the Comprehensive Test of Phonological Processing (CTOPP; Wagner, Torgesen, & Rashotte, 1999). On each trial, participants heard a list of phonemes or syllables and were asked to combine these elements into one pseudoword, or "nonword". For instance, if the participants heard /h/, /ɛ/, and /t/, they would

need to produce /hɛt/ as one word. The number of elements ranged from two to eight. Participants were given six practice trials and eighteen critical trials, and the dependent measure was the proportion of correct responses. Following the CTOPP procedure, responses were scored as either fully correct or incorrect, with no partial credit.

Phoneme Reversal. In the Phoneme Reversal task (CTOPP; Wagner et al., 1999), participants heard a pseudoword and were asked to repeat the word and then pronounce it backwards, creating a real English word. For instance, if the participants heard /stu:b/, they would need to produce the word *boots*. Participants were given four practice trials and eighteen critical trials, and the dependent measure was the proportion of correct responses. Following the CTOPP procedure, responses were scored as either fully correct or incorrect, with no partial credit.

Results

As we reviewed above, interpreting any relationship between self-paced reading times and the other constructs requires establishing that the measures are reliable (consistent). It is also critical to demonstrate that the measures are valid (measuring what they intend to measure). We thus first discuss the reliability and validity of, in turn, (a) the measures of verbal working memory, perceptual speed, inhibitory control, language experience, and phonological ability and (b) individual differences in syntactic processing in the self-paced reading task. Finally, we turn to whether any individual differences in syntactic processing—should we observe any—can be explained by the other cognitive constructs.

Individual Differences

Mean performance on all 16 individual difference measures across the five domains is summarized in Table 2.

The split-half correlations for each task are given in Table 3. These measures of internal consistency in individual differences on these tasks were generally on par with prior literature, indicating that we had successfully measured meaningful variation across individuals. However, the split-half correlations for Reading Span, NAART, Blending Nonwords, and Phoneme Reversal were noticeably lower than measures of internal consistency that have been reported in previous norms; this likely reflects the fact that our sample comprises a somewhat more restricted range of reading skills (Conway et al. 2005; Uttl, 2002; Wagner et al., 1997). The Eriksen flanker task had the lowest split-half correlations, which is perhaps not surprising given that the measure is a difference score and difference scores generally have lower consistency (e.g., Lord, 1963; Redick & Engle, 2006, but see Wostmann, Aichert, Costa, Rubia, Moller, and Ettinger, 2013, for higher consistency of the flanker test in other reports).

The next question was whether these individual differences reflect the underlying constructs that we expected them to. To assess this, we turned to the correlations between tasks. Ideally, tasks chosen to reflect the same underlying construct should exhibit moderately positive correlations, and tasks reflecting different constructs should be less correlated (e.g., Kane, Hambrick, Tuholski, Wilhelm, Payne, & Engle, 2004). Table 4 lists the correlations among all measures of individual differences of primary interest.

Additionally, in Table 5, we devote special attention (given past controversy on this point; e.g., MacDonald & Christiansen, 2002) to the correlations between the language experience composite measure and three aspects of the span measures: accuracy on the processing component of the task, the maximum time allotted to the processing component according to the calibration phase, and the actual span measure. Notably, language experience was significantly correlated with the response deadline set by the calibration phase in the

Reading Span task and with accuracy in the processing task in the Listening and Reading Span tasks, but it was crucially *not* correlated with the actual memory span scores for both Reading Span and Operation Span. This pattern implies that the calibration procedure was successful in separating the aspects of the Reading Span task that reflect linguistic experience (i.e., processing task speed and accuracy) from verbal working memory capacity per se. However, language experience was still correlated with Listening Span scores; This may be because the calibration procedure for Listening Span altered only the time taken to answer the questions and could not alter the presentation rate of the stimuli themselves.

In general, however, measurement properties of the individual differences tasks were mixed: Although many tasks behaved as expected, a few had internal consistency that was lower than expected, and the pattern of correlations among individual tasks did not align neatly with *a priori* constructs (i.e. some significant correlations between constructs and weaker correlations within constructs). As the inhibitory control battery was particularly problematic (consistent with the low convergent validity of this construct in other work; Duckworth & Kern, 2011), we also ran a version of the primary regression analyses below without these scores. Withholding inhibitory control had no effect on the overall pattern of results, except in one regression (see notes of Table 11). Therefore, we have chosen to retain all measures in the following analyses, but leave further exploration of these issues for future work.

Composite scores for each construct were devised by first standardizing all task scores (creating *z*-scores) and then averaging the standardized scores by subject within each domain (following Stine-Morrow et al., 2008). Table 6 lists the correlations among the composite scores from each domain. In general, the composite scores were correlated with one another. To assess whether this intercorrelation would be problematic for the regression analyses, we calculated

condition number as a measure of collinearity (Baayen, 2008; Belsley, Kuh, & Welsch, 1980). Among the five composite scores, condition number $\kappa = 1.68$, well within the range that indicates only weak collinearity (under 5; Belsley et al., 1980, p. 105).

Self-Paced Reading Measures

Prior to all other analyses, reading times were first corrected for word length by residualizing per-word reading times on word length (Ferreira & Clifton, 1986). Specifically, the dependent measure in the following reading time analyses was the residual of a linear regression model predicting log reading times from word length only (i.e., random slopes for subjects were not included; this was done to preserve subject-based variation for the individual differences analysis).

Residual reading times were then included as the dependent measure in a series of linear mixed-effects models to examine each of the syntactic phenomena of interest. Contrast coding was used for the sentence-type factors, producing main effect estimates comparable to those of an ANOVA. All models included random intercepts and random slopes for condition effects for subjects and items. Complete equations for the models can be found in Appendix A.

We first assessed whether we replicated the standard patterns in reading times across individuals (e.g., that verb bias interacts with ambiguity) for each of the three syntactic phenomena of interest. Then, we turn to whether we observed consistent individual differences in these phenomena, such that (for instance) some subjects consistently had larger verb bias effects than others.

Verb bias effects. For online processing of the verb-bias sentences, the critical region of analysis was defined as the embedded verb plus the following word (spillover). We constructed a

mixed-effects regression examining length-residualized reading times in the critical region as a function of ambiguity, verb bias, their interaction, and random effects for subjects and items.

The model yielded a significant ambiguity effect such that sentences without the complementizer *that* took longer to read ($\hat{\beta} = 0.074$, $SE = 0.0107$, $p < 0.001$), and it yielded a significant interaction with verb bias such that the ambiguity effect was larger for DO-biased sentences ($\hat{\beta} = 0.099$, $SE = 0.023$, $p < 0.001$). Together, these findings replicate the verb bias effect in the literature. Reading times across DO- and SC-biased sentences across both ambiguity conditions are plotted in Figure 1.

In offline comprehension, accuracy was high across all conditions (unambiguous SC = 92.8%; unambiguous DO = 92.0%; ambiguous SC= 93.1%; ambiguous DO = 92.0%; see Figure 2). As expected from the design of the comprehension questions, which probed general comprehension of the sentence rather than the DO/SC ambiguity specifically, there were no significant condition effects on accuracy (unambiguous: $\hat{\beta} = 0.050$, $SE = 0.115$, $p = 0.65$; SC-bias: $\hat{\beta} = -0.157$, $SE = 0.240$, $p = 0.511$; interaction: $\hat{\beta} = -0.212$, $SE = 0.180$, $p = 0.52$).

Extraction effects. Online reading times in the relative clause region were modeled as a function of relative clause condition (object- or subject-extracted) and random effects for subjects and items. Reading times were significantly longer in the ORC sentences relative to the SRC sentences ($\hat{\beta} = 0.222$, $SE = 0.050$, $p < 0.001$), replicating the standard finding in the literature. Reading times across both sentence types are plotted in Figure 3.

Offline comprehension accuracy across conditions is shown in Figure 4. Overall accuracy was highest when subjects were asked about the main clauses of SRC sentences (83.7%) and lowest when they were asked about the relative clauses of ORC sentences (71.5%); accuracy for the other trial types was closer to the high end (relative clause/SRC = 80.9%, main clause/ORC =

81.2%; see Figure 4). Accuracy on a given trial was modeled as a function of relative clause type, question type (whether than main clause or relative clause was probed), and random effects for subjects and items. Accuracy was significantly lower for ORC sentences ($\hat{\beta}$ = -0.446, *SE* = 0.143, $p < 0.01$). While the condition difference was numerically larger when the relative clause was probed rather than the main clause, this interaction was not significant ($\hat{\beta}$ = -0.386, *SE* = 0.286, $p = 0.18$).

Attachment preferences. As noted above, to distinguish participants' attachment preferences from any overall bias to affirm the readings provided in the comprehension questions, we used detection-theoretic analyses (Green & Swets, 1966; Macmillan & Creelman, 2004; Murayama et al., 2014; for applications to language processing, see Fraundorf, Watson, & Benjamin, 2010; Fraundorf, Benjamin, & Watson, 2013; Lee & Fraundorf, in press; Tokowicz & MacWhinney, 2005). In these models, the dependent variable is whether participants made a *yes* or *no* response to each comprehension question. Thus, a general response bias—across question types—to affirm the presented reading would be reflected in a significant intercept term whereas an effect of Condition (i.e., whether *yes* indicates a low attachment reading or a high attachment reading) on the odds of a *yes* response indicates whether participants preferred one type of attachment. Further, main effects of the individual-difference measures on the odds of a *yes* response would represent effects on the overall response bias whereas an interaction of the individual differences with the question type indicates effects on attachment preference.

The intercept term was significantly greater than 0 ($\hat{\beta}$ = 0.349, SE = 0.142, $p < .05$), indicating that participants did indeed have some overall preference to respond *yes*. However, this effect was small compared to significant main effect of question type: Participants gave far more *yes* responses when a *yes* response indicated a low attachment reading ($\hat{\beta}$ = 2.246, *SE* =

0.338, $p < 0.001$), consistent with prior results for attachment preferences in English. Figure 5 shows that subjects were more likely to endorse paraphrases consistent with the low attachment reading (75%) than the high attachment reading (36%).

Consistency of self-paced reading effects. The above analyses replicated the standard, across-participant effects from the language processing literature; for instance, ORCs were read more slowly than SRCs. However, critical for examining individual differences in syntactic processing is whether these effects consistently vary from subject to subject; that is, are there some subjects who consistently have more difficulty with ORCs than other subjects?

To assess the within-subject consistency of each effect, the data were randomly split into halves that were balanced on item and subject variables. Then, a regression model of the condition effects with random intercepts and slopes for subjects and items was run with each half of the data. Finally, the random effects for subjects from each of the two models were correlated as a measure of subject-level consistency. This entire process was repeated 100 times for each model, and the average correlation was taken as the final measure.

The results of this procedure are given in Table 7. The random intercepts for overall reading time showed a very high correlation between halves ($r = 0.96$ in verb bias sentences and $r = 0.95$ for relative-clause extraction sentences), indicating that *overall* individual differences in reading time differences were reliable. That is, some people were consistently faster readers than others, and we reliably measured this variability. Reliable differences in overall reading speed are expected and validate our analytical procedure as one that is capable of detecting individual differences that are known to exist. We also found that differences in overall comprehension accuracy were relatively consistent ($r = .57$ in verb-bias sentences and $r = .94$ in relative-clause extraction sentences).

But, the consistency of subject slopes for the syntactic variables (i.e., the individual differences in the syntactic-processing phenomena) was much lower, with all correlations of magnitude of .24 or lower for the online measures. Thus, for instance, while we replicated the *overall* verb bias effect, we did not observe consistent *individual differences* in the size of this effect. Correlations for the offline comprehension effects were of somewhat greater magnitude, but still relatively low.

Relation of Individual Differences to Language Comprehension

We did not observe strong evidence that some subjects consistently showed larger verb-bias effects than others. The pattern suggests that it should be difficult to observe relationships between those syntactic processing effects and the measures of individual differences since individual differences in syntactic processing either largely do not exist or could not be reliably measured.

Nevertheless, we considered on an exploratory basis whether individual differences in syntactic processing might be associated with the five other individual-difference constructs. In doing so, we are guided by the fact that relatively few other studies have assessed syntactic processing in conjunction with multiple other individual-difference constructs and across multiple syntactic phenomena; therefore, a gap in the literature could be filled by examining what relations might be observed in such data despite the limited reliability of the syntactic processing effects.

Reading times were examined as a simultaneous function of (a) syntactic condition variables for the relevant sentence type, (b) composite scores of all five cognitive domains measured in our individual differences battery, and (c) the interaction of the individual-difference scores with the syntactic condition variable(s). Individual differences that affect

syntactic processing (e.g., individual differences that differentially affect ORCs as opposed to SRCs) should be realized as an interaction between one of the individual-difference variables and syntactic condition. The purpose of including all five individual difference domains simultaneously in each regression was to allow us to interpret effects of one domain as accounting for a share of the variance independent of the other domains: As in other multiple regression models, parameter estimates in a mixed-effect regression reflect the effect of varying one variable (e.g., verbal working memory) while holding others constant (Baayen, 2008, p. 192). We included the composite individual-difference scores as continuous predictors to reflect the full range of these variables across individuals. Including continuous variation is more powerful than a median split (Cohen, 1983) and also yields more accurate estimates of effect size and lower rates of Type I error (MacCallum, Zhang, Preacher, & Rucker, 2002; Preacher, Rucker, MacCallum, & Nicewander, 2005). Appendix B presents the complete equations for these models.

For verb bias, although we replicated the *overall* verb bias effect in online reading (Table 8), we did not find that *individual differences* in the size of this effect were related to any of the other. Higher perceptual speed composite scores predicted faster reading times overall ($\hat{\beta}$ = -0.094, *SE* = 0.036, $p < 0.01$), but none of the individual difference measures significantly interacted with the syntactic condition effects. In *offline* performance (Table 9), there was a significant main effect of language experience, with higher scores leading to higher overall accuracy ($\hat{\beta}$ = 0.508, *SE* = 0.115, $p < 0.001$). Phonological ability also significantly interacted with ambiguity; it more strongly benefited accuracy in the ambiguous condition ($\hat{\beta}$ = -0.314, *SE* = 0.138, $p < 0.05$).

Similarly, for the extraction effects, although we replicated the overall difference between ORCs and SRCs in online reading (Table 10), the individual difference measures did not reveal any significant interactions with RC type. Again, however, there was an effect of perceptual speed on overall reading speed ($\hat{\beta}$ = -0.323, *SE* = 0.121, $p < 0.01$). There were also individual differences in offline comprehension (Table 11): Overall accuracy was significantly associated with higher scores in verbal working memory ($\hat{\beta}$ = 0.378, *SE* = 0.096, $p < 0.001$) and language experience ($\hat{\beta}$ = 0.437, *SE* = 0.123, $p < 0.001$). There was also a significant three-way interaction among language experience and the two condition effects; the difficulty of questions probing the object-extracted relative clauses was magnified for subjects with higher language experience scores ($\hat{\beta}$ = -0.653, *SE* = 0.314, $p < 0.05$).

Finally, for the attachment-preference items (Table 12), we did observe that several constructs related to individual differences in offline interpretation of these ambiguous items. Specifically, lower verbal working memory was associated with a stronger preference for high attachment ($\hat{\beta}$ = 1.020, *SE* = 0.292, $p < 0.001$), consistent with Swets et al. (2007) and Payne et al. (2014). Lower processing speed was also related to a high attachment preference ($\hat{\beta}$ = 0.656, *SE* = 0.255, $p < 0.05$). Additionally, both verbal working memory and language experience had significant effects on the overall response bias such that lower scores were associated with a higher *yes* bias (VWM: $\hat{\beta}$ = -0.471, *SE* = 0.156, $p < 0.01$; language experience: $\hat{\beta}$ = -0.415, *SE* = 0.192, $p < 0.05$).

Discussion

In summary, although we replicated across-participant online effects of verb bias and relative clause extraction type, individual differences in the magnitude of these effects were only seen offline. Verbal working memory and language experience were both shown to significantly

relate to overall offline accuracy for difference sentence types, and each also interacted with characteristics of the sentences: Language experience interacted with verb bias in offline comprehension of the verb bias sentences, while lower verbal working memory and slower perceptual speed were associated with a stronger high-attachment preference for RC attachment ambiguities. Below, we first discuss the absence of online effects, then the motivations for the reanalysis that is presented in Study 1b.

Effects of Individual Differences Are Offline, Not Online

Despite theories of sentence comprehension that predict online effects of verbal working memory, language experience, and other cognitive abilities, our study did not yield any interactions between the reading time effects and the individual differences assessed with our battery of tasks.

Why did we observe no relation between these individual differences and online syntactic processing? One possibility is that we simply did not measure the right individual-difference construct. However, *any* relation between an individual-difference construct and online syntactic processing would have actually been unlikely given that individual differences in syntactic processing had only moderate to low consistency to begin with. Thus, for instance, although the overall verb bias effect was robust across subjects, we did not observe strong evidence that some individuals consistently showed a larger verb bias effect than others. Because we did not observe strong individual differences in syntactic processing to begin with, no other construct could be expected to explain them.

It is important to emphasize that these limitations are specific to the consistency of individual differences in the syntactic phenomena of interest. We do not claim that the self-paced reading task fails to reliably measure reading time in general or even that it fails to

measure individual differences in reading time. In fact, as noted above, the split-half correlations for individual differences in overall reading speed, as assessed by subject-level random intercepts for reading time, were quite high (all *r*s > .9). Further, these differences in overall reading time correlated with individual differences in perceptual speed. Rather, it was specifically individual differences in the magnitude of the syntactic processing effects that were not consistent. Nor do we claim that the general (across-participant) syntactic processing phenomena of interest cannot be reliably observed. Indeed, when averaging across subjects, we replicated the standard findings from the literature (e.g., that ORCs are read more slowly than SRCs) for all of the sentence types presented here. That is, we observed both clear reader differences (differences in baseline reading speed and comprehension accuracy) and clear text differences (effects of verb bias and of relative clause extraction type). What we did not observe, at least at the level of syntactic processing, were consistent reader-text interactions whereby certain syntactic structures were particularly challenging for some readers.

In fact, the consistency across subjects of these syntactic-processing effects likely *contributes* to the absence of consistent individual differences. In general, effects that are robust and consistent across participants often make poorer individual-difference measures precisely because everyone exhibits the effect to similar degrees and there are few individual differences (see, for instance, Salthouse, Sieldlecki & Krueger, 2006, for similar results in memory control). For instance, let us return to the scenario mentioned above in which all subjects showed exactly a 300 ms reading-time difference between ORCs and SRCs. In this scenario, the consistency of the effect *across* subjects would make the overall effect highly robust and significant, but there would be no significant individual differences in the size of the extraction-type effect because the size of the effect does not differ across subjects. By contrast, if half of the subjects consistently

read ORCs more slowly than SRCs, and the other half consistently read SRCs more slowly than ORCs, individual differences are substantial, but there would be no significant across-participant effect because the individual differences average out to zero.

More broadly, the distinction between robust experimental effects and effective individual-difference measures reflects what are often differing goals in the two research traditions described by Cronbach (1957) and summarized in our introduction here: Experimental research most frequently seeks general principles of cognition that generalize across persons whereas individual-differences research generally seeks those characteristics that *differentiate* individuals. At the same time, experimental effects capitalize on the fact that a subject's task performance at a given time is subject to many *transient* influences, including the researchers' treatment variable(s) of interest. Therefore, the observed outcomes of interest in experimental settings are likely to be those that are most *malleable* within individuals, like reaction times.

In light of these principles, one possible explanation for why we did not observe consistent differences in online syntactic processing is that there are little or no individual differences in syntactic processing to begin with—that is, all readers find (for instance) ORCs more difficult than SRCs to similar degrees. Under this hypothesis, improvements in reliability of the measures would not reveal a relationship between online syntactic processing performance and other constructs because the underlying relationship does not exist. This claim would be consistent with the general framework put forth by Caplan and Waters (1999) suggesting a distinction between interpretive and post-interpretive processing, with span measures relating only to the latter. Caplan and Waters argued that interpretive processes, including word recognition and syntactic parsing, require a different resource pool than post-interpretive processes, such as encoding and reasoning about the input. This theory is supported by other

work that shows working memory span is unrelated to eye-tracking measures of differences between ORCs and SRCs in free reading (Traxler et al., 2005).

The other possibility, of course, is that there *are* individual differences in online syntactic processing, but the present study simply failed to measure them. Although we did find that the self-paced reading task reliably measured differences in overall reading speed, it is possible that features of the self-paced reading task itself may obscure individual differences in syntactic processing more specifically. Unlike in natural reading, the moving-window technique does not allow re-reading of prior material and requires readers to manually proceed through the sentence, making reading speed about twice as slow (Rayner, 1998, p. 391). These differences from natural reading may obscure typical individual tendencies in per-word reading times while still allowing variation in subsequent comprehension. Thus, it could be informative for future work to examine whether greater internal consistency in syntactic processing measures is obtained with eye-tracking of free reading. Some evidence suggests that *overall* subject-level differences in eye movements (e.g., individual differences in total reading time per word) can be reliably measured in free reading (Carter & Luke, 2016; Traxler et al., 2005). Further, at least one study (Traxler & Tooley, 2007) found that linguistic experience *did* correlate with individual differences in syntactic processing of DO/SC ambiguity items, unlike in the present study; this discrepancy might reflect the fact that Traxler and Tooley (2007) measured eye-tracking of free reading rather than the self-paced reading task.

Critically, because the observed *between-subject* variability in syntactic processing is met with substantial *within-subject* variability, we cannot fully disentangle these two possibilities. Regardless of its cause, we argue that the lack of *consistent* subject-level variation in syntactic processing condition effects in self-paced reading is unlikely to be exclusive to the present

dataset. The present task and materials were very similar to those used in other self-paced reading investigations of syntactic processing, suggesting that this lack of consistency is likely to extend to syntactic effects in self-paced reading time more broadly. In fact, low levels of internal consistency have been observed not just for online syntactic processing but for some other prospective individual differences derived from cognitive experiments, such as event-related potential measures of language comprehension (Tanner & Bulkes, 2015) and perspective-taking in comprehension (Brown-Schmidt & Fraundorf, 2015; Ryskin, Benjamin, Tullis, & Brown-Schmidt, 2015). The scope of these limitations is unclear: While standardized measures of domain-general cognitive ability and of linguistic experience have often been normed and show reliability, examining and reporting the consistency of individual differences in online language processing tasks themselves is less common (a problem that has also been noted elsewhere; Ryskin et al., 2015). It would be helpful for future investigations of individual differences in language processing to measure the internal consistency of individual differences in the online language processing measures themselves.

The Latent Variable Approach

Measurement issues make it difficult to draw clear conclusions from the current study. Thus, Study 1b provides a reanalysis of the data using a latent variable approach. In the following analyses and discussion, error in measurement is explicitly incorporated into the critical models. This is the topic of remainder of this chapter.

CHAPTER 3: REANALYSIS WITH LATENT VARIABLE APPROACH (STUDY 1B)

In the previous study, five constructs---language experience, phonological ability, working memory span, inhibitory control, and perceptual speed---were measured with a total of 16 tasks, where each construct was measured by at least two of these tasks. Five composite scores were then computed for each subject by standardizing and averaging the scores within each proposed construct. These scores were then entered as predictors in hierarchical multiple regression models for each of the five dependent measures.

The tasks were clustered into the five constructs a priori based upon previous uses of the tasks individually. Ideally, the observed scores would show evidence that the *a priori* clustering was justified. One way to assess whether this is the case is to observe the correlations between scores within each proposed construct and between proposed constructs. As reported in Table 4, the results are mixed.

Factor analysis is a formal approach to analyzing the covariance structure of a group of measurements, with the aim of summarizing these measurements as reflections of a smaller number of latent factors, or unobserved constructs. Specifically, a *confirmatory factor analysis* (CFA) is when the analyst has predictions about the number of underlying factors and about the relationship between the factors and the tests. These predictions are used to constrain the structure of the model; the model results are then compared with the observed correlations among the test scores to assess whether the predicted model structure is justified by the data.

Within this framework, the modeling of these latent factors as *predictors* of reading comprehension outcomes cane be accomplished through *structural equation modeling* (SEM). SEM is an inherently hierarchical generalization of multiple regression: at one level, the *structural model* estimates regression coefficients linking predictor latent variables (called

exogenous because their cause is not defined within the model) to dependent latent variables (called *endogenous* because their causes are defined within the model); at another level, the *measurement model* defines these latent variables in terms of the observed tasks scores that serve as *indicators*. In other words, the measurement model on its own is equivalent to a factor analytic model; SEM adds the structural model in the form of regression coefficients. Formally, the structural model is defined as

$$\boldsymbol{\eta} = \mathbf{B}\boldsymbol{\eta} + \boldsymbol{\Gamma}\boldsymbol{\xi} + \boldsymbol{\zeta}$$

where $\boldsymbol{\eta}$ is the $m \times 1$ vector of latent outcome variables, $\boldsymbol{\xi}$ is the $n \times 1$ vector of latent predictor variables, $\mathbf{B}$ is the $m \times m$ matrix of coefficients relating endogenous variables to one another (i.e. $\mathbf{B}$ is non-zero when at least one endogenous variable is hypothesized to predict another endogenous variable), $\boldsymbol{\Gamma}$ is the $m \times n$ matrix of coefficients relating exogenous variables to endogenous variables, and $\boldsymbol{\zeta}$ is the $m \times 1$ vector of latent residual error in predicting $\boldsymbol{\eta}$.

The measurement model is defined as

$$\mathbf{x} = \boldsymbol{\Lambda}_x \boldsymbol{\xi} + \boldsymbol{\delta}$$

$$\mathbf{y} = \boldsymbol{\Lambda}_y \boldsymbol{\eta} + \boldsymbol{\epsilon}$$

where $\mathbf{x}$ is the $q \times 1$ vector of indicator variables for $\boldsymbol{\xi}$, $\boldsymbol{\Lambda}_x$ is the $q \times n$ matrix of coefficients relating $\mathbf{x}$ to $\boldsymbol{\xi}$, and $\boldsymbol{\delta}$ is the $q \times 1$ vector of measurement errors for $\mathbf{x}$; analogously, $\mathbf{y}$ is the $p \times 1$ vector of indicator variables for $\boldsymbol{\eta}$, $\boldsymbol{\Lambda}_y$ is the $p \times m$ matrix of coefficients relating $\mathbf{y}$ to $\boldsymbol{\eta}$, and $\boldsymbol{\epsilon}$ is the $p \times 1$ vector of measurement errors for $\mathbf{y}$. A diagram of the full model is given in Figure X.

An advantage of this latent variable approach, relative to the approach in Study 1a of standardizing and combining scores, is that it allows measurement error to be modeled explicitly, rather than assuming that the combination of measurement errors from different tasks will "wash out" when combined. Issues with combining standardized scores are discussed further in Bollen

and Lennox (1991). More recently, Westfall and Yarkoni (2016) described the pitfalls of including unreliable predictors in regression models and recommend latent variable approaches as the proper alternative. In light of these issues, the aim of the current study is to perform a reanalysis of the data from Study 1a using CFA and SEM.

Analysis strategy

While SEM allows the measurement model and the structural model to be estimated simultaneously, a two-step approach of estimating the measurement model before the structural model is generally recommended over a one-step approach, as it makes it easier to diagnose causes of poor model fit (Anderson & Gerbing, 1988; Mueller & Hancock, 2008; Kline, 2016). Thus, the current analysis is broadly divided into two steps: first, measurement models of both the predictor and outcome variables are estimated with CFA; second, the measurement model suggested by the first analysis provides the bases for SEMs. Both steps take a somewhat exploratory approach: for the measurement model, an initial theory-driven model is estimated, followed by respecification that is guided both by theory and the current dataset; for the structural model, a full model with all latent predictors regressed onto the latent outcome is compared to critical models with fewer paths (driven by previous literature) and to a "no-path" model. A description of this model comparison approach is given in Anderson & Gerbing (1988) and an example with cognitive psychology data is given in Miyake, Friedman, Emerson, Witzki, Howerter, and Wager (2000).

Implementation

All of the analyses in the current study were conducted in R-Studio (Version 0.98.1103) using the `lavaan` package (Version 0.5-22; Rosseel, 2012). In all models, the variance in the latent variables was constrained to equal 1; errors were assumed to be uncorrelated; factor

loadings (**Λ**), covariances between latent predictor variables (**Φ** = $E(\xi\xi')$), and variances of the residuals ($\mathbf{\Theta_\delta} = E(\delta\delta')$; $\mathbf{\Theta_\epsilon} = E(\epsilon\epsilon')$) were freely estimated. Because the observed scores were standardized (for both exogenous and endogenous variables), intercepts for the observed variables were constrained to equal 0. Models were fit with full-information maximum likelihood estimation.

Data. In all models, the input data was a matrix of standardized observed scores. In the case of exogenous indicators (**x**), the input is the same as in Study 1a prior to creating composites within each construct. In the case of endogenous indicators (**y**), summary "scores" were calculated for each subject and each syntactic effect of interest. This is a departure from the MLM analysis in two ways: first, the data are now being averaged over trials, whereas the MLM analysis is predicting trial-level outcomes; second, in cases where the outcome of interest is a condition effect, the "score" for each subject is a *difference* in the average outcome across conditions (in the case of condition *interactions* of interest, such as the Verb Bias × Ambiguity effect, the score would be a difference of differences). As will be described in the following session, and as predicted from the generally poor internal consistency of the outcome variables discussed in Study 1a, the use of difference scores was problematic for model estimation.

If syntactic effects are the outcome variables of interest, and if these are considered to be *latent* variables that are measured with error (e.g. a subject has knowledge of verb biases, and we have measured it here as an observed difference in the ambiguity effect in direct-object- vs. sentential-complement-biased verb contexts), the next question is how to represent this latent variable in the SEM with only one experimental task per subject. Using the observed score as the single indicator for the latent construct does not account for measurement error, and thus does not take advantage of the latent variable approach. An alternative is to "parcel" the observed

score into subsets and use the subsets as indicators (Coffman & MacCallum, 2005; Williams & O'Boyle, 2008; Cole & Preacher, 2014). In the current analyses, trials in the self-paced reading task were split into balanced halves, difference scores were calculated in each half, and these two scores were each used as indicators for the latent syntactic effect of interest. In all cases, the parcels were constrained to have equal factor loadings.

Results

Measurement model of predictors

The first model was fit according to the assumptions of Study 1a, such that the 16 tasks measure five separable latent constructs, and that each task is a measure of only one construct. An initial version of this model used difference scores as the outcome measure for the Stroop and Flanker tasks, as in Study 1a. However, this resulted in non-significant loadings on the Inhibitory Control factor. Rather than exclude these measures or the Inhibitory Control factor, from the model, average reaction times in the conflict trials were used rather than different scores. Using conflict trial reaction times not only avoids the summation of measurement error that is inherent to difference scores, but it also aligns the outcome measures of these tasks to that of the third Inhibitory Control indicator, Anti-saccade (because there are no "pro-saccade" trials where the cue location predicts the target location, the outcome measure is only reaction times on "anti-saccade" or "conflict" trials). A diagram of this model is shown in Figure 7, and overall fit was fair (CFI = 0.917; SRMR = 0.088; RMSEA = 0.057, 90% CI = [0.032,0.078]).

Respecification. A clear problem in this model is that the loading from the Language Experience factor to the Reading Time Estimate (RTE) score is estimated to be about 0. This result is foreshadowed by the weak correlations between RTE and the other indicators of Language Experience, apart from Comparative Reading Habits (CRH), the other self-report

measure. Therefore, a second model structure was specified in which Language Experience was divided into two factors: Language Survey, indicated by RTE and CRH; and Language Skill, indicated by the remaining Language Experience tasks (NAART, ERVT, and ART). This second model is shown in Figure 8. A comparison of these two models indicated that the six-factor model was a better fit to the data ($\chi^2_{\text{diff}} = 19.238$, $df_{diff} = 5$, $p < 0.01$).

Although a close look at the local fit statistics (e.g. estimated residuals and modification indices) suggests that further adjustments to the model could be made in order to improve the fit (e.g. assigning indicators to different groupings), the six-factor model described above was retained because it remained consistent with existing theory, preserved unidimensionality of the indicators, and provided an acceptable overall fit (CFI = 0.947; SRMR = 0.081; RMSEA = 0.047, 90% CI = [0.013,0.070]). Therefore, this first set of analyses has resulted in two substantive changes from Study 1a that will carry over into the upcoming SEM analyses: first, difference scores were replaced with reaction times from the theoretically more difficult conflict trials; second, the model was restructured such that the Language Experience factor was divided into Language Skill and Language Survey factors.

Measurement model of outcome variables

The second component of the measurement model is the relationship between the latent outcomes and their indicators. As explained above, the observed outcome measures were divided into parcels such that each latent outcome had two indicators. Although the SEM analyses will treat each outcome separately such that each model has only one latent outcome (**η**), latent outcomes of interest were modeled simultaneously here because a single-factor model with less than three indicators is not identifiable (Bollen, 1989, p. 244). Technically, the following models are CFAs just as above, meaning that the latent variables are exogenous in this context because

their hypothesized causes have not yet entered the model. However, I will continue to refer to these as "outcomes" to remain consistent with the upcoming SEM analyses.

Overall outcomes. The first set of outcomes indicate overall performance: (1) mean reading times and (2) mean accuracy across Verb Bias and Relative Clause Extraction experimental sentences, and (3) overall "yes" bias in the Relative Clause Attachment experimental sentences. Coefficients between these outcomes and the predictor variables in the upcoming SEM analyses would be analogous to main effects of individual differences variables in the Study 1a analyses. The measurement model for the overall outcomes is shown in Figure 9. The results indicate a good model fit (CFI = 0.998; SRMR = 0.022; RMSEA = 0.026, 90% CI = [0.000,0.103]).

Syntactic effects. The critical set of outcomes are the syntactic effects: (1) the verb bias effect on reading times, such that reading times should be longest in the direct object (DO) - biased and Ambiguous condition (a Verb Bias × Ambiguity interaction); (2) the relative clause extraction effect on reading times such that object-extracted relative clauses (ORCs) should take longer to read that subject-extracted relative clauses (SRCs); (3) the relative clause effect on offline comprehension such that accuracy should be particularly poor when the relative-clause region is probed following an ORC (a Question Type × Clause Type interaction); and (4) relative clause attachment preference, realized here as an effect of question type (low- or high-attachment endorsement) on the tendency to respond "yes", such that a positive difference between low- and high-attachment conditions indicates a low-attachment preference. The verb bias condition effects on accuracy were excluded from this analysis because no such condition effect was predicted, and thus the relationship between individual differences and condition effects are not of interest here. Condition effects were implemented in the model as difference scores (outcomes

(2) and (4), above); interaction effects were implemented in the model as difference-of-difference scores (outcomes (1) and (3), above).

An initial model was specified as described and failed to converge. As discussed previously, the low internal consistency for some of these outcomes suggested that using difference scores as indicators would be problematic. Accordingly, a second model was fit in which all difference scores, excluding that which indicated attachment preference, were exchanged for average performance in the theoretically most difficult condition (i.e. reading times in the DO-biased/Ambiguous condition, reading times in the ORC condition, and accuracy in the ORC/Relative-Clause Question condition for outcomes (1), (2), and (3), respectively). The diagram for this model is given in Figure 10. Overall, this model provided a good fit to the data (CFI = 0.984; SRMR = 0.064; RMSEA = 0.057, 90% CI = [0.000,0.103]).

Implications. As a result of these two sets of analyses, estimating measurement models for the exogenous and endogenous latent variables of interest, the following SEMs will include six exogenous factors, each with at least two indicators, and a single endogenous factor that is indicated by two "parcels". The general form of this model is shown in Figure 11.

Structural equation models

For each outcome of interest, a series of nested models were fit and compared. A full or "all-path" model was fit such that regression coefficients toward the outcome were freely estimated for all six latent predictor variables, analogous to the critical regression models in Study 1a. This was compared to a null or "no-path" model in which regression paths are fixed at 0, as are the covariances between the latent endogenous variable and all each of the six exogenous factors. Finally, for the models of syntactic effects, an intermediate "theory-driven" model was fit such that only a subset of the regression coefficients are estimated, representing

the factors with the strongest theoretical support. The specification of these theory-driven models is admittedly subjective and is mainly for exploratory purposes, increasing the number of candidate models.

Overall outcomes. For each of the three overall outcomes—reading time, accuracy, and "yes" bias—the full model was preferred over the null model (such that $p\,(>\chi^2_{\text{diff}}) < 0.01$) and the full model demonstrated acceptable fit. The regression coefficients and fit statistics for these models are given in Table 13. Interestingly, the pattern of significant regression coefficients differs from the Study 1a results. First, although main effects of language experience were found in predicting accuracy and "yes" bias, the analogous regression coefficients were not significant here. Second, overall reading time is now predicted by Inhibitory Control and not Perceptual Speed as in Study 1a, although it is important to note that the indicators for Inhibitory Control are now conflated with overall reaction time differences.

Syntactic effects. The critical models in the current study are those predicting syntactic effects: online effects of verb bias, online and offline effects of relative clause extraction, and relative clause attachment preferences. Each of these four models will be described in turn.

***Verb bias and reading times*.** In these models, the observed dependent measure was reading time for DO-biased/Ambiguous sentences. The full and no-path models were compared to a model with both the Language Skill and Language Survey paths (e.g. MacDonald et al., 1994) and the Inhibitory Control path (e.g. Novick, Trueswell, & Thompson-Schill, 2010; Christianson, Hollingworth, Halliwell, & Ferreira, 2001). The full model indicates that Inhibitory Control is the only significant predictor of verb bias ($\gamma = 0.485$, SE = 0.187, $p < 0.001$). Model comparisons endorsed the reduced model, as it was not significantly different from the full model ($\chi^2_{\text{diff}} = 0.820$, $df_{diff} = 3$, $p = 0.845$) and it is more parsimonious. The reduced model

replicates the significant Inhibitory Control Path, as well as a very small but statistically significant path coefficient for the Language Survey path. Results of these models and their comparisons are given in Table 14.

Relative clause extraction and reading times. In these models, the observed dependent measure was average reading time for ORC sentences. The full and no-path models were compared to a model with Verbal Working Memory (e.g., Gibson, 1998, 2000; Traxler, 2007, Swets et al., 2007), and both Language paths (e.g. MacDonald & Christiansen, 2002). Neither the full model nor the reduced model yield any significant regression coefficients. Although the full model was favored over the reduced model ($p < 0.05$), it only marginally outperformed the no-path model (which in turn, was not significantly different than the reduced model). This pattern of seemingly contradictory model comparisons, paired with the non-significant path coefficients, suggests that the best choice of model is the no-path model, as it is most parsimonious. Results of these models and their comparisons are given in Table 15.

Relative clause extraction and accuracy. In these models, the observed dependent measure was the overall accuracy in answering comprehension questions in the ORC/Relative Clause Question condition. The full and no-path models were compared to a model with Verbal Working Memory, Language Skill, and Language Survey paths, just as in the case of reading times. In the full model, all path coefficients are significant, aside from Language Skill. In the reduced model, only the Verbal Working Memory path coefficient is significant. The full model is favored over the other two in model comparisons. Results of these models and their comparisons are given in Table 16.

Attachment preferences. Finally, the last set of models predicted relative clause attachment preferences, indicated by the difference in "yes" responses between the low- and

high-attachment endorsing questions. Here, the full and no-path models were compared to a model with Verbal Working Memory (e.g. Gibson, 2000; Swets et al., 2007, Payne et al., 2014) and both Language paths (Payne et al., 2014). The full model yielded significant path coefficients for Verbal Working Memory ($\gamma = 0.535$, SE = 0.181, $p < 0.01$) and Perceptual Speed ($\gamma = 0.511$, SE = 0.228, $p < 0.05$), such that higher values of both were related to a low attachment preference. This is consistent with the previous analysis in Study 1a. The reduced model again yielded a significant coefficient for Verbal Working Memory ($\gamma = 0.488$, SE = 0.145, $p < 0.001$), and a very small but statistically significant coefficient for the Language Survey factor ($\gamma = 0.008$, SE = 0.003, $p < 0.01$). Model comparisons endorsed the reduced model, which was significantly better than the no-path model ($\chi^2_{\text{diff}} = 27.397$, $df_{diff} = 3$, $p < 0.001$), and only marginally worse than the full model ($\chi^2_{\text{diff}} = 7.036$, $df_{diff} = 3$, $p = 0.071$). Results of these models and their comparisons are given in Table 17.

Discussion

In summary, Study 1b presented a reanalysis of the data in Study 1a, in which 133 participants read experimentally manipulated sentences and completed a battery of tasks designed to capture relevant individual differences. The reanalysis took a latent variable approach, allowing the measurement concerns highlighted in Study 1a to be re-examined and explicitly accounted for in the critical regression analyses. In some areas, the reanalysis provided additional support to conclusions from Study 1a. Namely, that (1) aside from the condition effect that indicates relative clause attachment preferences, the indicators for syntactic processing effects have low reliability when used as indicators of *individuals'* processing; and (2) high attachment is associated with lower working memory and slower perceptual speed, in line with results from Swets et al. (2007) and Payne et al. (2014).

Crucially, Study 1b also provides new results that contribute to the initial conclusions of Study 1a. First, the measurement models provide more information about the relationships among the observed variables and their relative strengths as indicators. Second, measurement model comparisons provided an explicit way to test alternative conceptions of "language experience" as an individual trait. The better fit of the six-factor model suggests that future work would benefit from subdividing the multifaceted construct into more specific components, each with multiple indicators chosen *a priori*.

Unsurprisingly, in the models predicting outcomes with low internal consistency, and in models involving modified observed variables, there are discrepancies between the two analytic approaches in which factors are significant predictors. The two approaches also have several important differences that need to be considered. While Study 1b has the advantage of explicitly modeling measurement error, which can lead to spurious regression results when left unchecked (see e.g. Westfall & Yarkoni, 2016), there are also some limitations of this second set of analyses. First, the latent variable analyses were entirely *post hoc*, and thus the study design and resulting data were not optimized for SEM. Ideally, each latent variable would have more than two indicators: "Two *might* be fine, three is better, four is best, and anything more is gravy" (Kenny, 1979, p. 143, as quoted by Mueller & Hancock, 2008). SEM analyses would also benefit from a larger sample of participants; although sample size recommendations vary (see Kline, 2016 for discussion), SEM is a "large-sample technique" with increasing numbers of subjects needed for models with many observed variables, for structurally-complex models, for estimation techniques that are robust to non-normally distributed data, for observed scores with low reliability, or performing cross-validation (Thompson, 2000, p. 272; Kline, 2016, pp. 14-16).

Another limitation in the current latent variable analyses is that the outcome variables were submitted to the model after averaging over items within subjects, both for the exogenous variables (as was done in Study 1a's analyses), and the endogenous variables. In part, this is an issue of power; fitting a model in which items serve as indicators for performance on tasks, which in turn serve as indicators for latent constructs, would increase the complexity of the model and dramatically increase the number of parameters to estimate. There is also the issue of how (or *whether*) to derive measures of subject-level traits from performance on experimental tasks. Here, in the case of endogenous variables representing syntactic effects, the initial choice to use condition differences resulted in unstable models, leading to the data from only one experimental cell being used in each analysis. Avoiding difference scores reduces measurement error, but it also changes the interpretation of the measure, conflating baseline difference in speed or accuracy with the effects of syntactic complexity. This issue is not unique to the current set of studies (e.g. Stroop and Flanker are examples of experimental tasks that are used as measures of stable individual differences in the broader literature, and that suffer from similar measurement issues; see Duckworth & Kern, 2011). This highlights a major challenge in uniting the experimental and correlational disciplines: the accurate summary of an individual subject's sensitivity to an experimental effect requires that *both* the baseline condition and experimental condition(s) are accurate measures of the subject's performance, leading to an informative comparison across these conditions.

Finally, in both sets of analyses, it would be ideal to validate the predicted models on an independent sample of participants to assess the generalizability of the model solutions given here.

CHAPTER 4: CONCLUSIONS

Taken together, Studies 1a and 1b demonstrate that there are considerable individual differences in both sentence comprehension performance and in language-specific and domain-general abilities within a young, college-educated adults, and that there is some evidence for consistent relationships between comprehension and these more general abilities. Testing the influences of multiple factors on three "test cases" in sentence processing literature within a single set of subjects, this work provides an unprecedented opportunity to examine key theoretical mechanisms in the psycholinguistic literature.

Critically, this work also brings to light issues of measurement reliability and validity within subjects, which are often overlooked in experimental studies, even when individual differences are of theoretical interest. Generally, measures derived from experimental tasks (i.e. measures of inhibitory control and reading outcomes) were not well-suited to serve as indicators for individuals' latent abilities, as evidenced by the poor internal consistency of within-person *condition effects*. While the latent variable analysis in Study 1b did yield a significant relationship between a general cognitive ability (Inhibitory Control) and verb bias, this was only after the measures were converted from difference scores into average reading or reaction times within a single condition. Thus, baseline and experimental effects are conflated, and the interpretation of this significant regression coefficient is made unclear.

Verbal Working Memory Capacity

However, one effect that is well supported by the current study is the tendency for high relative clause attachment to be preferred by individuals with lower working memory and slower perceptual speed. These results obtained across both sets of analyses and are consistent with Swets et al. (2007) and Payne et al. (2014). Swets et al. (2007) present evidence that this relation

obtains because individuals with limited memory resources adopt a chunking strategy that generates an implicit prosodic representation amenable to a high-attachment reading, analogous to the effects of explicit prosody (for a review of those effects, see Frazier, Carlson, & Clifton, 2006). These results contradict predictions that people with lower memory resources would prefer the low-attachment reading because it minimizes distance of the dependency (e.g., Gibson, 1998, 2000; Traxler, 2007). Rather, this finding supports the account that internal prosody may be an important strategy for readers with low working memory (see Swets et al., 2007).

Lower verbal working memory was also associated with lower overall accuracy on RC extraction sentences, consistent with other findings that verbal working memory relates to general reading comprehension accuracy (Daneman & Merikle, 1996, for meta-analysis). Note, however, that this effect of working memory was a main effect across all of the RC extraction sentences and did not differentially affect ORCs relative to SRCs. Thus, it can be best characterized as a *reader* effect—lower-span readers have poorer comprehension—rather than a reader-text interaction whereby lower-span readers are particularly disadvantaged with specific types of relative clauses.

Nevertheless, it is noteworthy that an effect of verbal working memory on reading comprehension emerged even when other factors, including language experience, were included in the model. It is worth noting that our span tasks included a calibration phase intended to help deconfound language experience and span, and the results presented in Table 5 indicate that this procedure was generally successful. It is possible that this step allowed us to observe an independent contribution of verbal working memory span. Nevertheless, it should be noted that all the span measures were specific to *verbal* working memory; we can make no claims about

how syntactic processing may be affected by any other possible forms of working memory (e.g, visuospatial working memory; Shah & Miyake, 1996).

Language Experience

In the first set of analyses, higher scores on the language experience measures were indeed related to higher *overall* comprehension accuracy on both verb bias and RC extraction sentences, consistent with several studies demonstrating that individuals with more exposure to language have higher reading comprehension skills (e.g., Stanovich, 1985). However, in the offline measures, there was only inconsistent evidence for the language experience × syntax interactions that would indicate language comprehension facilitated comprehension of specific syntactic structures, and we did not observe effects of language experience at all on self-paced reading times. As mentioned previously, the inconsistency of individual differences in the online syntactic effects may make such relations difficult to detect. Further, it is unclear that more *total* language experience would necessarily improve processing of the dispreferred structure. Rather, it may be the *relative* exposure to these more difficult structures that is crucial. For instance, in the training studies by Wells et al. (2009) and Fine et al. (2013), comprehension of the uncommon structures may have improved with additional exposures because those specific structures were proportionally more frequent in the training input than in the typical distributional statistics participants had experienced previously. In fact, increased overall exposure to the typical distributional statistics of English may only strengthen biases against the statistically dispreferred structure, such as the high-attachment processing cost observed for high-print-exposure older adults in Payne and colleagues' (2014) study.

Another possibility is that the differences in exposure within our educated adult sample are not great enough to modulate online effects. However, this possibility is unlikely because

scores on the individual differences tasks, including the language experience tasks, were well distributed across the range of possible scores rather than being clustered at the ceiling (see Table 1). Further, the correlations across tasks are evidence against a ceiling effect; the fact that some subjects consistently score higher than others across tasks suggests that not all subjects are at ceiling, and that there are individual differences even within this restricted range. Nonetheless, including a wider range of prior language experience should be a goal of future research.

Summary

Each experience of reading brings together both a reader and a to-be-read text. Research in educational and differential psychology has made it clear that some people are more skilled readers than others, and psycholinguists have revealed that some syntactic structures are more difficult than others (e.g., subject-extracted relative clauses are more difficult than object-extracted relative clauses). What has been less clear, at least in the domain of syntactic processing, is whether there are reader-text interactions: Are there particular syntactic structures that are especially challenging for particular readers, and are there some readers who are especially advantaged at reading otherwise difficult structures?

We investigated whether such interactions exist in syntactic processing and, if so, what other individual differences might drive them. In doing so, we were guided by several insights from the correlational approach: We measured multiple individual-difference constructs, we obtained multiple measures of each construct, and we assessed the consistency of our measures. Our results suggest a possible reason for the lack of consensus across studies that have examined individual differences in syntactic processing: the relatively low consistency of those differences, especially in online (rather than offline) measures. Although we replicated well-studied syntactic phenomena *overall* (e.g., the effects of verb distributional statistics), we found low

consistency of *individual differences* in those effects: That is, it was *not* the case that some subjects consistently showed (for instance) large verb bias effects and other subjects consistently showed small verb bias effects.

By contrast, we did observe both reliable differences among individuals in their overall reading speed and in baseline comprehension accuracy (i.e., reader effects). We also observed that some syntactic structures were more challenging than others (i.e., text effects). What we found little evidence of—within the domain of syntactic processing—were reader-text interactions whereby some syntactic structures were differentially more difficult for some readers than others. Rather, good readers were comparatively good with all syntactic structures, and syntactically challenging sentences were more difficult for all readers, and these two effects did not interact. However, reader-text interactions have been observed for some aspects of language processing *other* than syntactic processing, such as lexical or discourse-level processing (e.g., Seidenberg, 1985; Stine-Morrow et al., 2008).

Our results are thus consistent with psycholinguistic theories in which initial stages of syntactic processing are relatively automatic and not influenced by domain-general processes (Caplan & Waters, 1999). And, they imply that the skills and abilities that educational psychologists have identified as underlying reading success are likely relevant across a broad range of syntactic structures. Consequently, it may not be necessary to tailor texts or interventions to particular readers, at least at the level of syntax. Nevertheless, it will be important to bring the language-processing effects demonstrated by experimental psychologists together with the important individual differences identified by correlational approach in order to identify how they might or might not interact—and to progress towards a united discipline of correlational and experimental traditions in psychology.

TABLES

Table 1

Inter-rater reliability for tasks requiring subjective scoring.

Task	$N_{subjects}$	$N_{ratings}$	Match proportion	Cohen's κ
NAART	100	6100	0.884	0.765
Stroop accuracy	36	7200	0.980	0.816
Pseudoword Repetition	12	1056	0.863	0.848
Blending Nonwords	96	2304	0.954	0.900
Phoneme Reversal	107	2354	0.992	0.981

Note. $N_{subjects}$ is the number of subjects that were scored by two raters, and $N_{ratings}$ is the total number of trials with two ratings for these subjects. *Match proportion* is the proportion of $N_{ratings}$ that were the same across both raters. Cohen's κ provides a correction for chance agreement among raters (Cohen, 1960).

Table 2

Summary of task performance on measures of individual differences.

Construct	Task	Measure	Min.	Mean	Max.	*SD*
Language experience	ART	Number correct with penalty	-9	10	47	11.570
	ERVT		1	17.180	36.750	7.740
	CRH	Sum of Likert responses	5	22	33	5.326
	NAART	Number correct	0.283	0.560	0.885	0.125
	RTE	Hours per week	5	20	63	10.695
Verbal working memory span	Listening	Score with partial credit	5.633	8.943	10	0.954
	Operation		1.852	10.588	15	3.435
	Reading		2.367	6.749	10	1.757
Inhibitory control	Anti. Acc.	Proportion correct responses (all conflict trials)	0.264	0.717	0.986	0.192
	Anti. RT	Log median reaction time for correct responses (all conflict trials)	-2.043	-0.559	0.270	0.328
	Flanker	Difference in log median reaction time of correct responses for conflict and no-conflict trials	-0.254	0.151	0.344	0.069
	Stroop		-0.153	0.193	0.514	0.135
Phonological ability	Pseudo. Rep.	Proportion correct	0.487	0.801	0.949	0.077
	BNW		0.167	0.646	1	0.176
	PR		0.182	0.687	1	0.176
Perceptual speed	Letter	Number correct within time limit	41	73	405	34.305
	Pattern		46	84	398	31.378

Notes: Min. and *Max.* refer to the observed minimum and maximum scores, respectively. ART = Author Recognition Test (Stanovich & West, 1989; Acheson, Wells, & MacDonald, 2008); ERVT = Extended Range Vocabulary Test (Ekstrom, French, Harman, & Dermen, 1976); CRH = Comparative Reading Habits (Acheson et al., 2008); NAART = North American Adult Reading Test (Blair & Spreen, 1989); RTE = Reading Time Estimate (Acheson et al., 2008); "Listening" = Listening Span task (Daneman & Carpenter, 1986; Stine & Hindman, 1994; Unsworth, Heitz, Schrock, & Engle, 2005); "Operation" = Operation Span task (Turner & Engle, 1989; Unsworth et al., 2005); "Reading" = Reading Span task (Daneman & Carpenter, 1986; Stine & Hindman, 1994; Unsworth, Heitz, Schrock, & Engle, 2005); "Anti.Acc" and "Anti.RT" refer to accuracy and median reaction time on the Antisaccade Task (Kane, Bleckley, Conway, & Engle, 2001), respectively; "Flanker" = Flanker task (Eriksen & Eriksen, 1974); ; "Stroop" = Stroop task (Stroop, 1935; Brown-Schmidt, 2009); "Pseudo.Rep" = Pseudoword Repetition (Gupta, 2003); BNW = Blending Nonwords (Wagner, Torgesen, & Rashotte, 1999); PR = Phoneme Reversal (Wagner, Torgesen, & Rashotte, 1999); "Letter" = Letter Comparison (Salthouse & Babcock, 1991); "Pattern" = Pattern Comparison (Salthouse & Babcock, 1991).

Table 3

Split half correlations of individual differences tasks

Construct	Task	Split half correlation
Language experience	ART	0.721
	ERVT	0.702
	NAART	0.827
Verbal working memory	Reading	0.623
	Listening	0.574
	Operation	0.807
Inhibitory control	Flanker	0.287
	Anti. Acc.	0.891
	Anti. RT	0.892
	Stroop	0.834
Phonological ability	Pseudo. Rep.	0.791
	BNW	0.531
	PR	0.459
Perceptual speed	Letter	0.893
	Pattern	0.916

Notes: Tasks were split into balanced halves, scores were calculated for each subject in each half in the same manner as in the main analyses, and the two lists of subject scores were correlated. ART = Author Recognition Test (Stanovich & West, 1989; Acheson, Wells, & MacDonald, 2008); ERVT = Extended Range Vocabulary Test (Ekstrom, French, Harman, & Dermen, 1976); CRH = Comparative Reading Habits (Acheson et al., 2008); NAART = North American Adult Reading Test (Blair & Spreen, 1989); RTE = Reading Time Estimate (Acheson et al., 2008); "Listening" = Listening Span task (Daneman & Carpenter, 1986; Stine & Hindman, 1994; Unsworth, Heitz, Schrock, & Engle, 2005); "Operation" = Operation Span task (Turner & Engle, 1989; Unsworth et al., 2005); "Reading" = Reading Span task (Daneman & Carpenter, 1986; Stine & Hindman, 1994; Unsworth, Heitz, Schrock, & Engle, 2005); "Anti.Acc" and "Anti.RT" refer to accuracy and median reaction time on the Antisaccade Task (Kane, Bleckley, Conway, & Engle, 2001), respectively; "Flanker" = Flanker task (Eriksen & Eriksen, 1974); ; "Stroop" = Stroop task (Stroop, 1935; Brown-Schmidt, 2009); "Pseudo.Rep" = Pseudoword Repetition (Gupta, 2003); BNW = Blending Nonwords (Wagner, Torgesen, & Rashotte, 1999); PR = Phoneme Reversal (Wagner, Torgesen, & Rashotte, 1999); "Letter" = Letter Comparison (Salthouse & Babcock, 1991); "Pattern" = Pattern Comparison (Salthouse & Babcock, 1991).

Table 4

Correlations among measures of individual differences

Speed		Language experience					Verbal Working Memory			
LComp	PComp	NAART	RTE	CRH	Vocab	ART	OSpan	LSpan	RSpan	
-0.027	0.015	0.184	-0.137	0.166	0.063	-0.036	0.511***	0.398***	1.000	RSpan
0.176	0.177	0.181	-0.001	0.180	0.204*	0.131	0.493***	1.000	0.398***	LSpan
0.076	0.128	0.025	-0.181	0.069	0.060	0.050	1.000	0.493***	0.511***	OSpan
-0.038	-0.047	0.386***	0.194*	0.248**	0.453***	1.000	0.050	0.131	-0.036	ART
0.191*	0.202*	0.682***	0.104	0.349***	1.000	0.453***	0.060	0.204*	0.063	Vocab
0.019	0.016	0.263**	0.387***	1.000	0.349***	0.248**	0.069	0.180	0.166	CRH
-0.106	-0.039	0.182	1.000	0.387***	0.104	0.194*	-0.181	-0.001	-0.137	RTE
0.221*	0.211*	1.000	0.182	0.263**	0.682***	0.386***	0.025	0.181	0.184	NAART
0.682***	1.000	0.211*	-0.039	0.016	0.202*	-0.047	0.128	0.177	0.015	PComp
1.000	0.682***	0.221*	-0.106	0.019	0.191*	-0.038	0.076	0.176	-0.027	LComp
0.101	0.175	0.504***	0.010	0.201*	0.451***	0.277**	0.142	0.148	0.197*	Gupta
0.084	0.139	0.336***	-0.079	0.153	0.198*	0.028	0.032	0.069	0.166	BNW
0.137	0.186*	0.299**	-0.167	-0.044	0.329***	0.223*	0.172	0.129	0.199*	PR
-0.182*	-0.110	-0.225*	0.123	0.007	-0.099	-0.081	-0.197	-0.269**	0.027	Stroop
0.049	0.163	0.055	0.068	-0.042	0.112	-0.103	-0.020	0.027	0.130	Flanker
0.217*	0.328***	0.161	-0.333***	0.051	0.264**	-0.061	0.319***	0.207*	0.301***	Anti.Acc

Table 4, cont.

	Inhibitory Control			Phon. Ability		
	Anti.Acc	Flanker	Stroop	PR	BNW	Gupta
RSpan	0.301***	0.130	0.027	0.199*	0.166	0.197*
LSpan	0.207*	0.027	-0.269**	0.129	0.069	0.148
OSpan	0.319***	-0.020	-0.197	0.172	0.032	0.142
ART	-0.061	-0.103	-0.081	0.223*	0.028	0.277**
Vocab	0.264**	0.112	-0.099	0.329***	0.198*	0.451***
CRH	0.051	-0.042	0.007	-0.044	0.153	0.201*
RTE	-0.333***	0.068	0.123	-0.167	-0.079	0.010
NAART	0.161	0.055	-0.225*	0.299**	0.336***	0.504***
PComp	0.328***	0.163	-0.110	0.186*	0.139	0.175
LComp	0.217*	0.049	-0.182*	0.137	0.084	0.101
Gupta	0.233*	0.096	-0.169	0.322***	0.459***	1.000
BNW	0.290**	0.128	-0.104	0.345***	1.000	0.459***
PR	0.401***	0.107	-0.135	1.000	0.345***	0.322***
Stroop	-0.204*	-0.049	1.000	-0.135	-0.104	-0.169
Flanker	0.258**	1.000	-0.049	0.107	0.128	0.096
Anti.Acc	1.000	0.258**	-0.204*	0.401***	0.290**	0.233*

Notes: *** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$.

Table 5

Correlations among components of memory span tasks and language composite score

	Processing accuracy		Calibrated time limit		Span score	
Task	Pearson's *r*	*t*-value	Pearson's *r*	*t*-value	Pearson's *r*	*t*-value
Reading	0.397***	4.899	-0.237**	-2.762	0.092	1.027
Listening	0.312***	3.715	-0.142	-1.618	0.214*	2.375
Operation	0.178*	2.039	0.012	0.134	0.054	0.558

Note: *** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$

Table 6

Correlations among composite scores

	Lang.	WM	Inhib.	Speed
VWM	0.161			
Inhib.	0.078	0.327***		
Speed	0.134	0.088	0.241**	
Phon.	0.333***	0.299***	0.299***	0.194*

Notes: *** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$. "VWM" = Verbal working memory span; "Inhib" = Inhibitory control; "Speed" = Perceptual speed; "Phon." = Phonological ability; "Lang." = Language experience.

Table 7

Mean correlation of random subject effects in split halves

Model	Random effect	*r*	95% CI
Verb bias reading time	Intercept	0.956	[0.954, 0.958]
	Ambig	-0.051	[-0.089, -0.012]
	Bias	-0.103	[-0.133, -0.073]
	Interaction	0.237	[0.211, 0.263]
Verb bias accuracy	Intercept	0.573	[0.565, 0.582]
	Ambig	-0.176	[-0.223, -0.13]
	Bias	0.426	[0.382, 0.471]
	Interaction	0.216	[0.152, 0.281]
RC-extraction reading time	Intercept	0.948	[0.947, 0.95]
	Slope	0.053	[0.016, 0.09]
RC-extraction accuracy	Intercept	0.941	[0.941, 0.941]
	RC	-0.005	[-0.084, 0.074]
	Question	0.229	[0.163, 0.295]
	Interaction	0.198	[0.13, 0.265]
Attachment preference	Intercept	0.033	[-0.031, 0.096]
	Slope	0.678	[0.67, 0.685]

Note: Correlations were calculated by running each of the five models of condition effects on randomly-generated halves of the data and correlating the random effects. Means were calculated by repeating the correlation procedure 100 times and averaging over the results. The 95% confidence interval (CI) is also given.

Table 8

Fixed effects in model of residual reading times in Verb Bias sentences

	Fixed effect	Estimate	*SE*	*p*-value
	(Intercept)	0.042	0.042	0.015
Individual differences	VWM	0.040	0.040	0.938
	Inhib.	0.057	0.057	0.375
	Lang.	0.048	0.048	0.835
	Phon.	0.048	0.048	0.971
	Speed	0.036	0.036	0.010
Condition effects	Ambiguous	0.016	0.016	0.000
	DO Bias	0.057	0.057	0.567
	Ambiguous x DO Bias	0.031	0.031	0.003
Individual difference x Condition effect interactions	VWM x Ambiguous	0.014	0.014	0.710
	VWM x DO Bias	0.014	0.014	0.298
	Inhib. x Ambiguous	0.021	0.021	0.959
	Inhib. x DO Bias	0.020	0.020	0.575
	Lang. x Ambiguous	0.017	0.017	0.783
	Lang. x DO Bias	0.017	0.017	0.815
	Phon. x Ambiguous	0.018	0.018	0.200
	Phon. x DO Bias	0.017	0.017	0.332
	Speed x Ambiguous	0.013	0.013	0.910
	Speed x DO Bias	0.013	0.013	0.095
	VWM x Ambiguous x DO Bias	0.029	0.029	0.509
	Inhib. x Ambiguous x DO Bias	0.041	0.041	0.573
	Lang. x Ambiguous x DO Bias	0.034	0.034	0.518
	Phon. x Ambiguous x DO Bias	0.035	0.035	0.564
	Speed x Ambiguous x DO Bias	0.026	0.026	0.089

Notes: Contrast coding was used for condition effects. Condition effects here refer to the change in reading time when sentences were ambiguous (opposed to unambiguous) and DO-biased (opposed to SC-biased). Random intercepts and slopes for all condition effects for both subjects and items were also included in the model. "VWM" = Verbal working memory span; "Inhib" = Inhibitory control; "Speed" = Perceptual speed; "Phon." = Phonological ability; "Lang." = Language experience.

Table 9

Fixed effects in model of comprehension accuracy in Verb Bias sentences

	Fixed effect	Estimate	*SE*	*p-value*
	(Intercept)	3.274	0.142	<0.001
Individual differences	VWM	0.052	0.087	0.552
	Inhib.	0.223	0.125	0.075
	Phon.	0.086	0.106	0.416
	Lang.	0.508	0.115	<0.001
	Speed	0.000	0.078	0.999
Condition effects	Unambiguous	0.109	0.128	0.392
	SC Bias	-0.231	0.261	0.377
	Unambiguous x SC Bias	-0.238	0.201	0.236
Individual difference x Condition effect interactions	VWM x Unambiguous	0.057	0.113	0.614
	VWM x SC Bias	-0.057	0.110	0.601
	Inhib. x Unambiguous	-0.092	0.163	0.574
	Inhib. x SC Bias	-0.137	0.159	0.388
	Phon. x Unambiguous	-0.314	0.138	0.023
	Phon. x SC Bias	0.104	0.134	0.438
	Lang. x Unambiguous	0.312	0.162	0.054
	Lang. x SC Bias	-0.270	0.159	0.089
	Speed x Unambiguous	-0.016	0.101	0.875
	Speed x SC Bias	0.059	0.098	0.545
	VWM x Unambiguous x SC Bias	0.165	0.219	0.453
	Inhib. x Unambiguous x SC Bias	-0.076	0.318	0.812
	Phon. x Unambiguous x SC Bias	-0.059	0.269	0.826
	Lang. x Unambiguous x SC Bias	-0.555	0.318	0.081
	Speed x Unambiguous x SC Bias	-0.347	0.197	0.079

Notes: Contrast coding was used for condition effects. Condition effects here refer to the change in log odds of a correct response when the sentences were unambiguous (opposed to ambiguous) and SC-biased (opposed to DO-biased). Random intercepts and slopes for all condition effects for both subjects and items were also included in the model. "VWM" = Verbal working memory span; "Inhib" = Inhibitory control; "Speed" = Perceptual speed; "Phon." = Phonological ability; "Lang." = Language experience.

Table 10

Fixed effects in model of residual reading times in Relative Clause Extraction sentences

	Fixed effect	Estimate	*SE*	*p*-value
	(Intercept)	-0.222	0.164	0.179
Individual differences	VWM	0.069	0.135	0.611
	Inhib.	-0.106	0.194	0.586
	Phon.	0.070	0.164	0.671
	Lang.	-0.099	0.162	0.542
	Speed	-0.323	0.121	0.009
Condition effect	ORC	0.217	0.056	<0.001
Individual difference x Condition interactions	VWM x ORC	0.000	0.062	0.994
	Inhib. x ORC	-0.117	0.088	0.189
	Phon. x ORC	0.031	0.075	0.683
	Lang. x ORC	-0.114	0.073	0.121
	Speed x ORC	0.075	0.055	0.173

Notes: Contrast coding was used for the condition effect. The condition effect here refers to the change in reading time when given an object-extracted relative clause (opposed to subject-extracted). Random intercepts and slopes for all condition effects for both subjects and items were also included in the model. "VWM" = Verbal orking memory span; "Inhib" = Inhibitory control; "Speed" = Perceptual speed; "Phon." = Phonological ability; "Lang." = Language experience.

Table 11

Fixed effects in model of comprehension accuracy in Relative Clause Extraction sentences

	Fixed effect	Estimate	*SE*	*p*-value
	(Intercept)	1.850	0.186	<0.001
Individual differences	VWM	0.378	0.096	<0.001
	Inhib.	-0.025	0.140	0.860
	Phon.	0.194	0.120	0.104
	Lang.	0.437	0.123	<0.001
	Speed	0.103	0.089	0.246
Condition effects	ORC	-0.464	0.155	0.003
	RCQ	-0.595	0.353	0.092
	ORC x RCQ	-0.472	0.309	0.127
Individual difference x Condition interactions	VWM x ORC	0.102	0.115	0.376
	VWM x RCQ	0.058	0.113	0.605
	Inhib. x ORC	0.147	0.169	0.386
	Inhib. x RCQ	-0.245	0.166	0.139
	Phon. x ORC	-0.180	0.147	0.221
	Phon. x RCQ[a]	-0.249	0.143	0.082
	Lang. x ORC	-0.020	0.158	0.898
	Lang. x RCQ	-0.190	0.155	0.222
	Speed x ORC[a]	0.195	0.109	0.074
	Speed x RCQ	0.091	0.107	0.395
	VWM x ORC x RCQ	-0.013	0.227	0.955
	Inhib. x ORC x RCQ	-0.012	0.333	0.971
	Phon. x ORC x RCQ	0.363	0.289	0.209
	Lang. x ORC x RCQ	-0.653	0.314	0.037
	Speed x ORC x RCQ	-0.325	0.216	0.131

Notes: Contrast coding was used for the condition effects. The condition effects here refers to the change in the log odds of correct responding when given an object-extracted relative clause (opposed to subject-extracted) and receiving a comprehension question that probed the relative clause region of the sentences (opposed to the main clause). Random intercepts and slopes for all condition effects for both subjects and items were also included in the model. "VWM" = Verbal working memory span; "Inhib" = Inhibitory control; "Speed" = Perceptual speed; "Phon." = Phonological ability; "Lang." = Language experience. [a] These effects reached significance in the model if inhibitory control and its interactions were dropped.

Table 12

Fixed effects in model of relative clause attachment preferences

	Fixed effect	Estimate	*SE*	*p*-value
	(Intercept)	-0.892	0.244	<0.001
Condition effect	Low Attach.	2.465	0.344	<0.001
Individual differences	VWM	-0.471	0.156	0.003
	Inhib.	-0.013	0.225	0.953
	Phon.	-0.128	0.191	0.501
	Lang.	-0.415	0.192	0.030
	Speed	-0.277	0.144	0.053
Individual difference x Condition interactions	Low Attach. x VWM	0.938	0.275	0.001
	Low Attach. x Inhib.	-0.045	0.399	0.910
	Low Attach. x Phon.	0.251	0.338	0.457
	Low Attach. x Lang.	0.339	0.335	0.311
	Low Attach. x Speed	0.656	0.255	0.010

Notes: Contrast coding was used for the condition effect. The condition effect here refers to the likelihood of answering *yes* to a comprehension question when it promoted the low attachment reading (opposed to the high attachment reading). Random intercepts and slopes for all condition effects for both subjects and items were also included in the model. "VWM" = Verbal working memory span; "Inhib" = Inhibitory control; "Speed" = Perceptual speed; "Phon." = Phonological ability; "Lang." = Language experience.

Table 13

Models predicting overall reading outcomes

Model	df	χ^2	$p(\chi^2)$	SRMR	RMSEA [90%CI]	CFI	AIC	Path coefficients (SE)					
								Lang Skill	Lang Survey	vWMC	Speed	Inhib	Phon
Overall accuracy	133	145.614	0.215	0.079	0.027 [0, 0.051]	0.979	5724.382	0.454 (0.303)	0.252 (0.238)	0.642 (0.230)**	0.467 (0.253).	0.575 (0.030)*	0.009 (0.335)
Overall reading time	133	143.278	0.256	0.077	0.024 [0, 0.050]	0.988	5505.395	-0.276 (0.242)	0.133 (0.158)	-0.161 (0.140)	-0.070 (0.175)	0.449 (0.187)	0.230 (0.253)
Overall "yes" bias	133	144.694	0.230	0.079	0.026 [0, 0.051]	0.977	5830.109	-0.710 (0.444)	-0.006 (0.005)	0.002 (0.256)	-0.281 (0.343)	-0.625 (0.407)	0.139 (0.460)

Table 14

Models predicting individual verb bias effects on reading times

Model	df	χ^2	$p(\chi^2)$	SRMR	RMSEA [90%CI]	CFI	AIC	Path coefficients (SE)						Comparison	
								Lang Skill	Lang Survey	vWMC	Speed	Inhib	Phon	Model 2	Model 3
Model 1: All paths	133	137.746	0.371	0.078	0.017 [0, 0.046]	0.993	5674.514	-0.074 (0.195)	0.009 (0.007)	-0.120 (0.134)	-0.062 (0.174)	0.485 (0.187)***	0.0101 (0.236)	p = 0.820	p < 0.001
Model 2: Lang. paths	136	138.566	0.423	0.079	0.012 [0, 0.044]	0.996	5669.334	-0.050 (0.104)	0.006 (0.002)*	--	--	0.558 (0.125)***	--	--	p < 0.001
Model 3: No paths	139	169.104	0.042	0.105	0.041 [0, 0.061]	0.954	5693.871	--	--	--	--	--	--	--	--

Table 15

Models predicting individual relative clause extraction effects on reading times

Model	df	χ^2	$p(\chi^2)$	SRMR	RMSEA [90%CI]	CFI	AIC	Path coefficients (SE)						Comparison	
								Lang Skill	Lang Survey	vWMC	Speed	Inhib	Phon	Model 2	Model 3
Model 1: All paths	133	146.668	0.197	0.078	0.028 [0, 0.052]	0.976	5751.242	-0.366 (0.288)	0.061 (0.180)	-0.053 (0.154)	-0.232 (0.187)	0.079 (0.192)	0.357 (0.305)	$p <$ 0.05	$p < 0.1$
Model 2: Lang. and VWM paths	136	155.553	0.120	0.085	0.033 [0, 0.055]	0.966	5754.128	-0.164 (0.132)	0.049 (0.134)	0.006 (0.120)	--	--	--	--	$p =$ 0.620
Model 3: No paths	139	157.331	0.137	0.087	0.032 [0, 0.054]	0.968	5749.905	--	--	--	--	--	--	--	--

Table 16

Models predicting individual relative clause extraction x question effects on accuracy

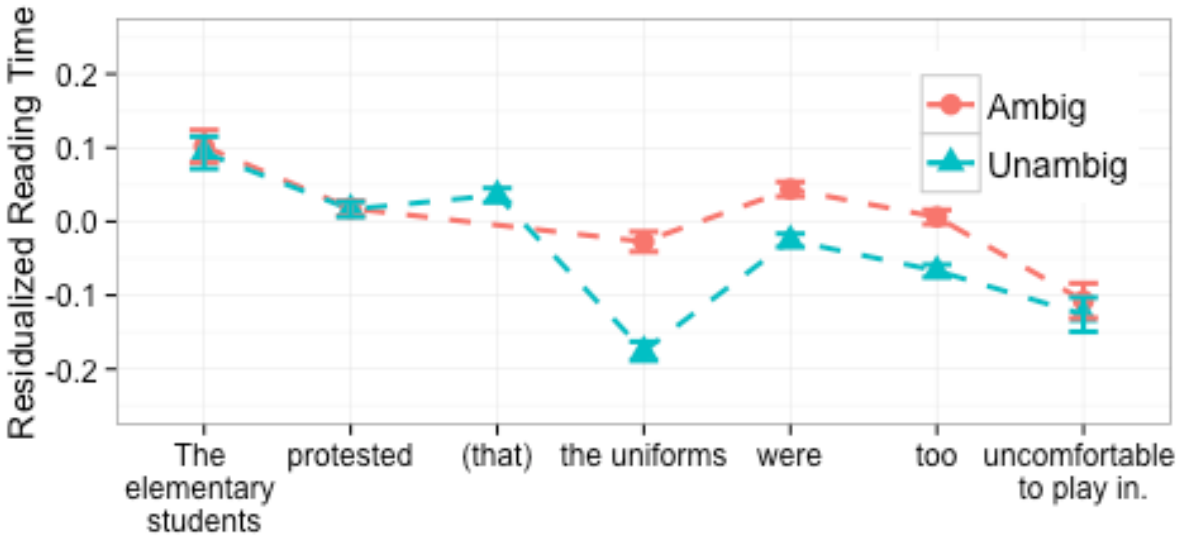
Model	df	χ^2	$p(\chi^2)$	SRMR	RMSEA [90%CI]	CFI	AIC	Path coefficients (SE)						Comparison	
								Lang Skill	Lang Survey	vWMC	Speed	Inhib	Phon	Model 2	Model 3
Model 1: All paths	133	142.539	0.270	0.081	0.023 [0, 0.049]	0.981	5817.964	48.735 (63.924)	111.901 (48.312)*	560.355 (9.228)***	477.904 (57.532)***	422.685 (61.734)***	-170.331 (49.226)***	$p <$ 0.05	$p <$ 0.001
Model 2: Lang. and vWM paths	136	151.855	0.167	0.083	0.030 [0, 0.053]	0.969	5821.281	32.030 (60.295)	52.835 (56.619)	224.420 (12.488)***	--	--	--	--	$p <$ 0.001
Model 3: No paths	139	179.299	0.012	0.100	0.047 [0.023, 0.066]	0.921	5842.724	--	--	--	--	--	--	--	--

Table 17

Models predicting individual relative clause low attachment bias

Model	df	χ^2	$p(\chi^2)$	SRMR	RMSEA [90%CI]	CFI	AIC	Path coefficients (SE)						Comparison	
								Lang Skill	Lang Survey	vWMC	Speed	Inhib	Phon	Model 2	Model 3
Model 1: All paths*	133	136.013	0.411	0.076	0.013 [0, 0.045]	0.995	5730.931	0.020 (0.240)	0.075 (0.389)	0.535 (0.181) **	0.511 (0.228) *	0.411 (0.231) .	-0.068 (0.282)	p = 0.071	p < 0.001
Model 2: Lang. and vWM paths*	136	143.063	0.322	0.079	0.020 [0, 0.047]	0.988	5731.967	0.134 (0.123)	0.008 (0.003) **	0.488 (0.145) ***	--	--	--		p < 0.001
Model 3: No paths	139	170.460	0.036	0.102	0.042 [0, 0.061]	0.947	5753.363	--	--	--	--	--	--		

a.



b.

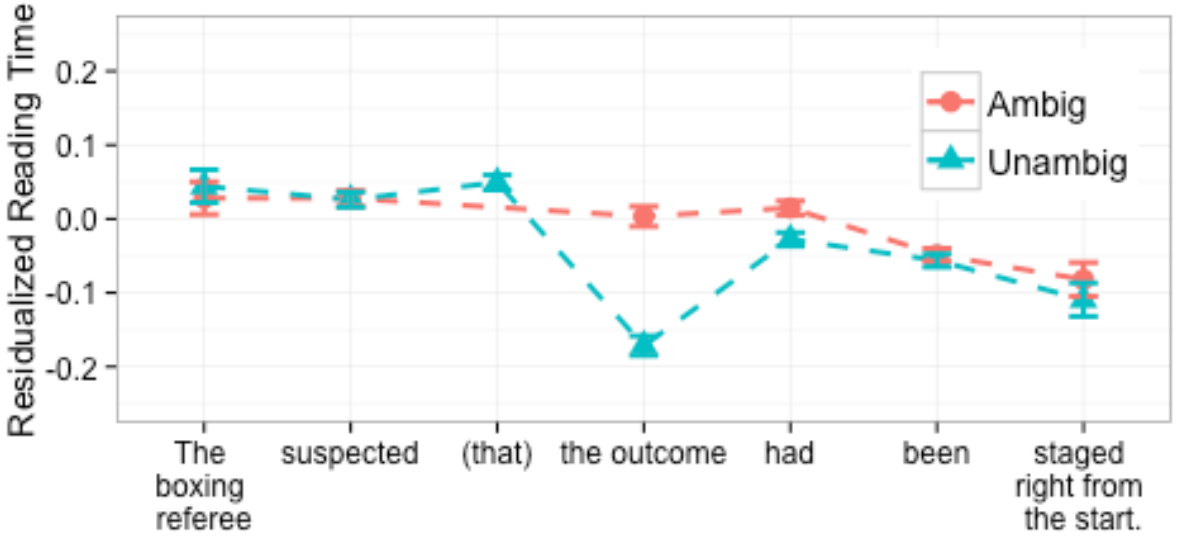


Figure 1. Mean residual reading times across sentence regions for DO-biased (a) and SC-biased (b) sentences. Error bars represent the standard errors of the mean residual reading times with the correction for within-subject factors given in Morey (2008).

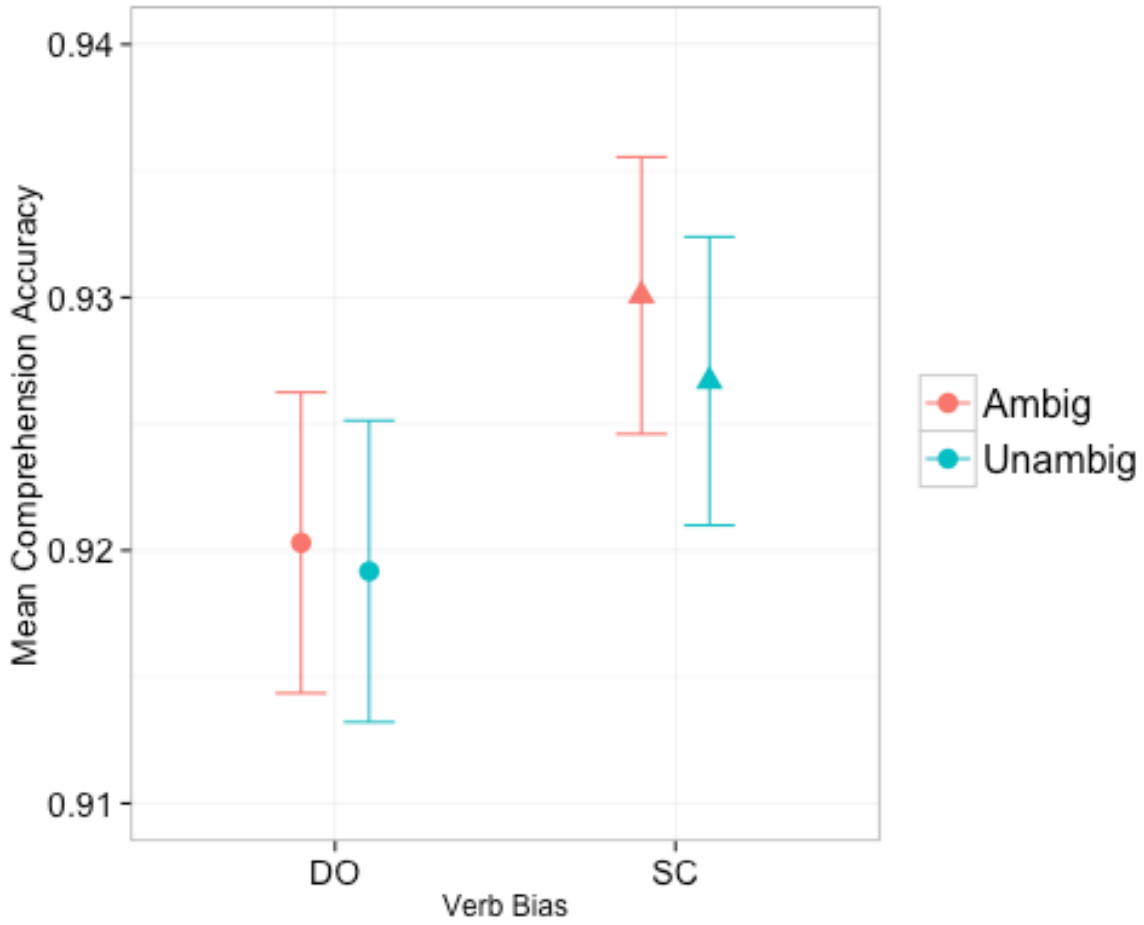


Figure 2. Mean comprehension question accuracy on ambiguous and unambiguous direct object (DO)- and sentential complement (SC)-biased sentences. Error bars represent the standard errors of the mean residual reading times with the correction for within-subject factors given in Morey (2008).

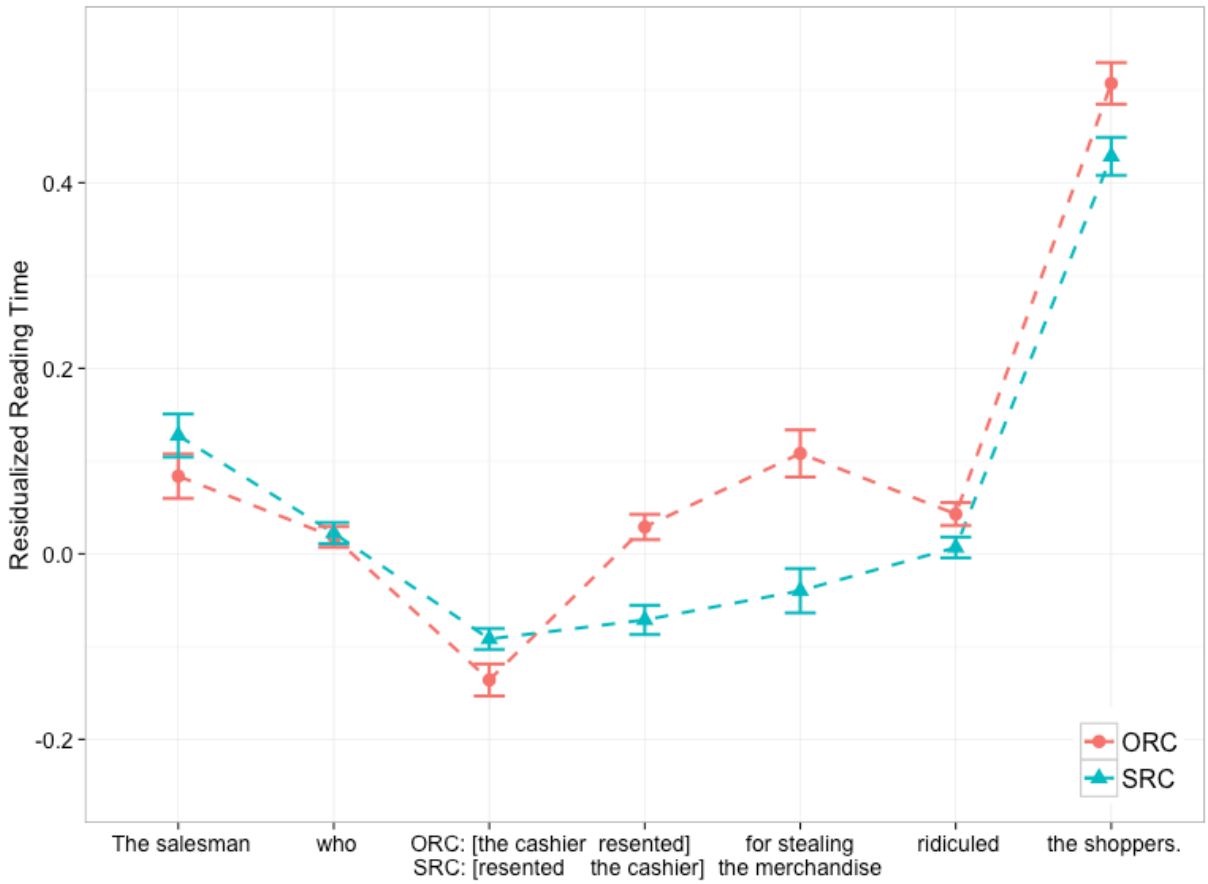


Figure 3. Mean residual reading times across sentence regions for object- (ORC) and subject-extracted (SRC) relative clause sentences. Error bars represent the standard errors of the mean residual reading times with the correction for within-subject factors given in Morey (2008).

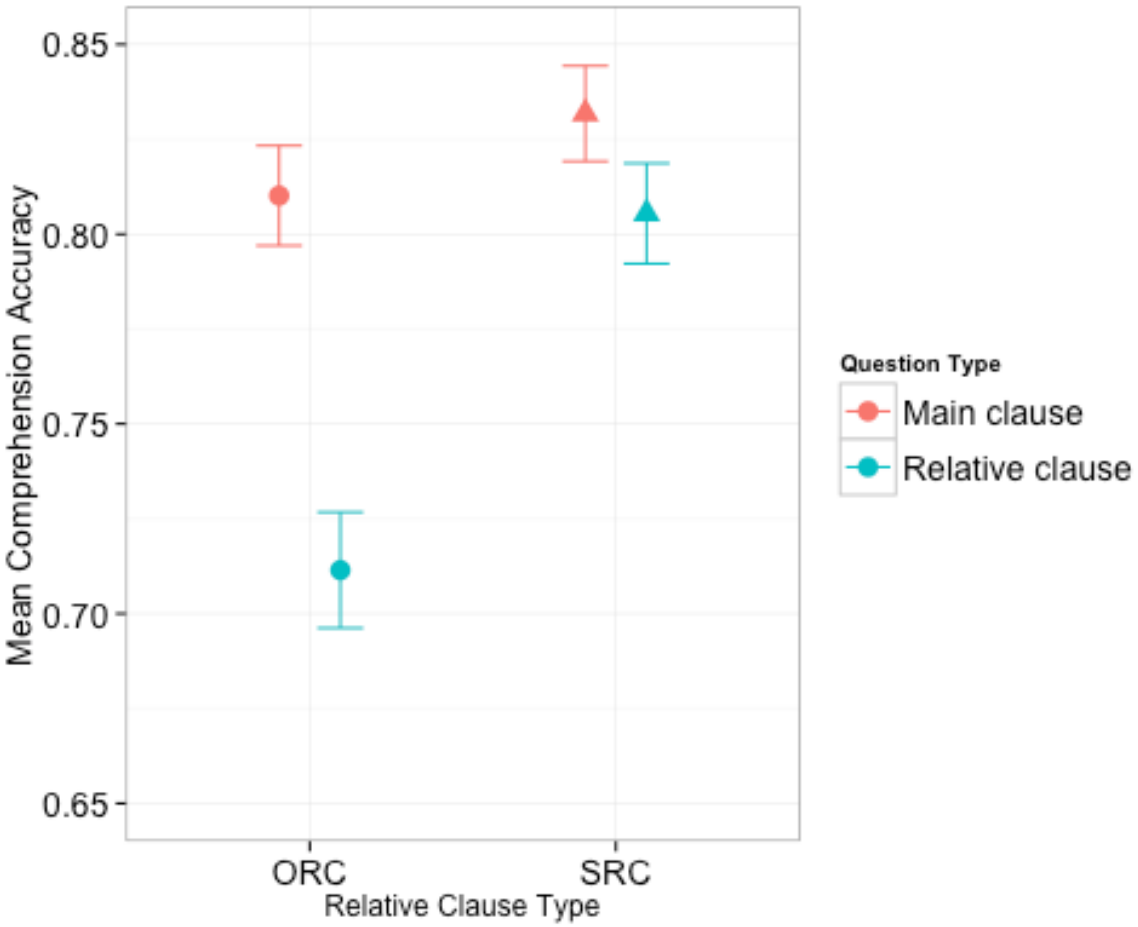


Figure 4. Mean comprehension question accuracy for object- (ORC) and subject-extracted (SRC) relative clause sentences, by question type (whether the main clause or the relative clause was probed). Error bars represent the standard errors of the mean residual reading times with the correction for within-subject factors given in Morey (2008).

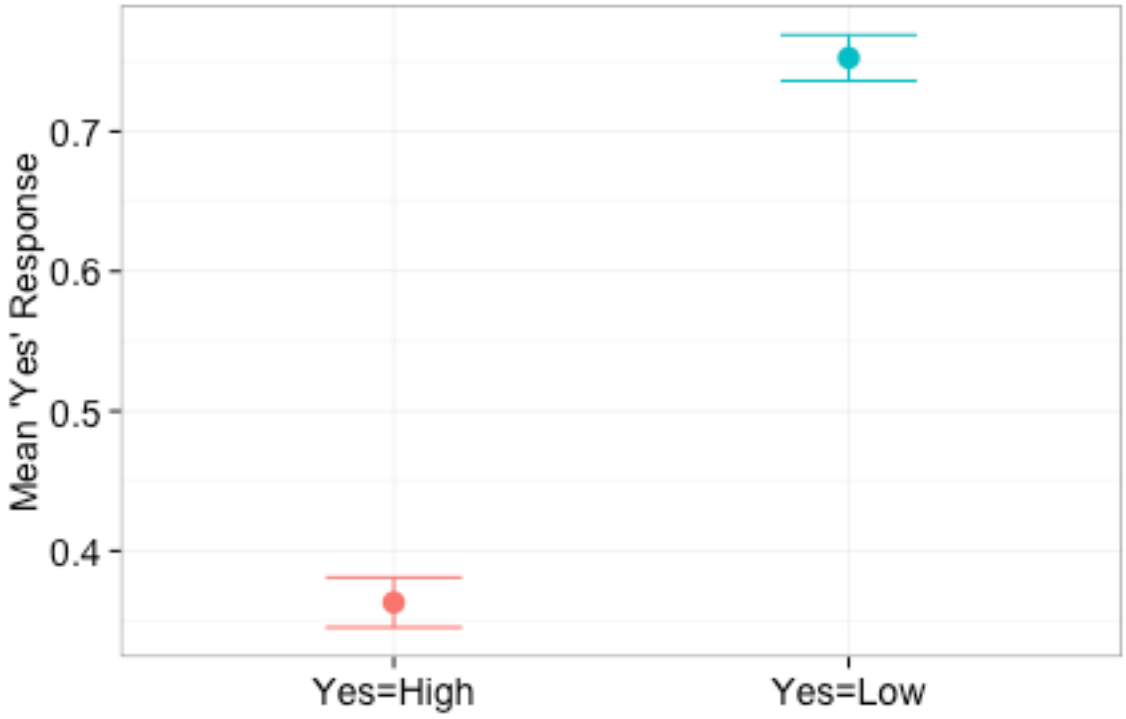


Figure 5. Mean proportion of *yes* responses to comprehension questions after sentences containing a global relative clause attachment ambiguity, according to whether the question suggested a high attachment reading ("Yes=High") or a low attachment reading ("Yes=Low"). Error bars represent the standard errors of the mean residual reading times with the correction for within-subject factors given in Morey (2008).

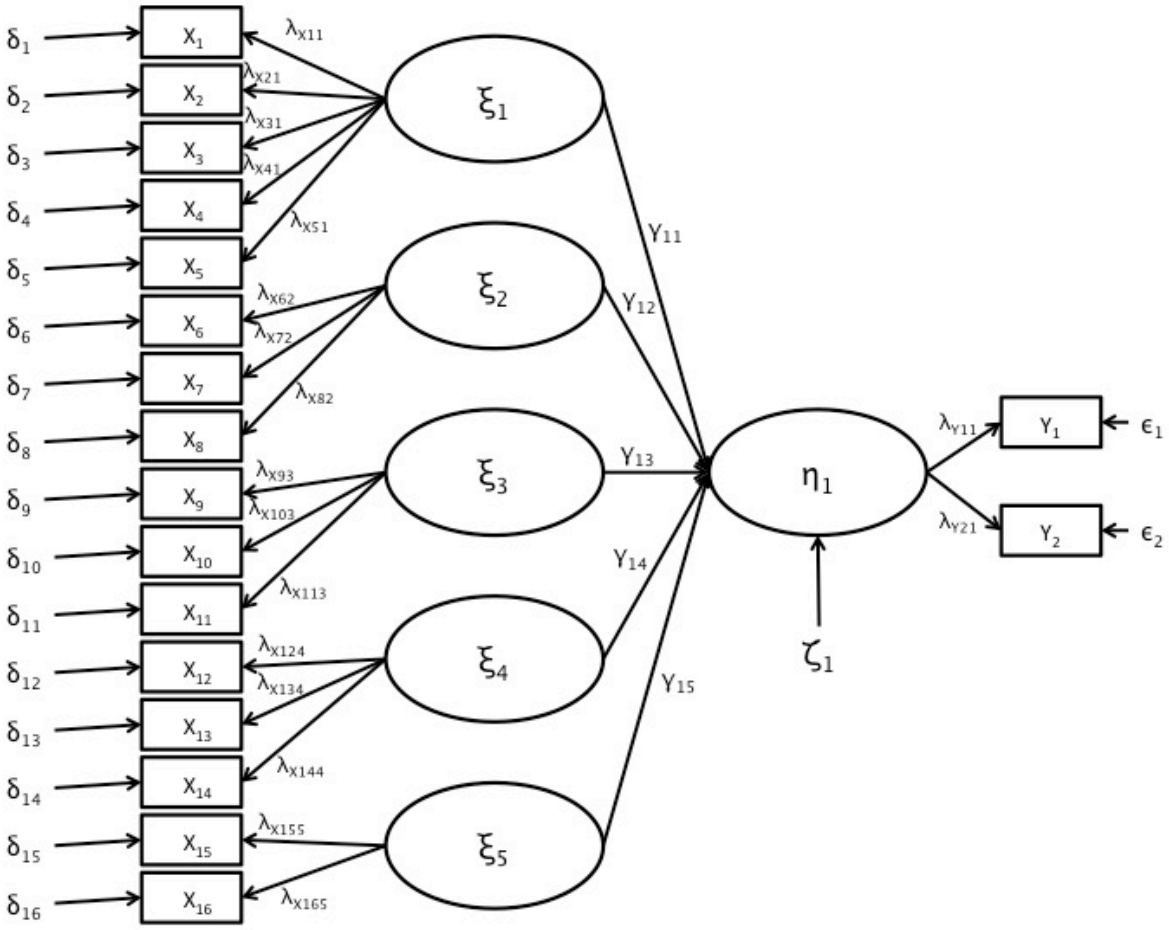


Figure 6. Structural equation model (SEM) diagram with general notation.

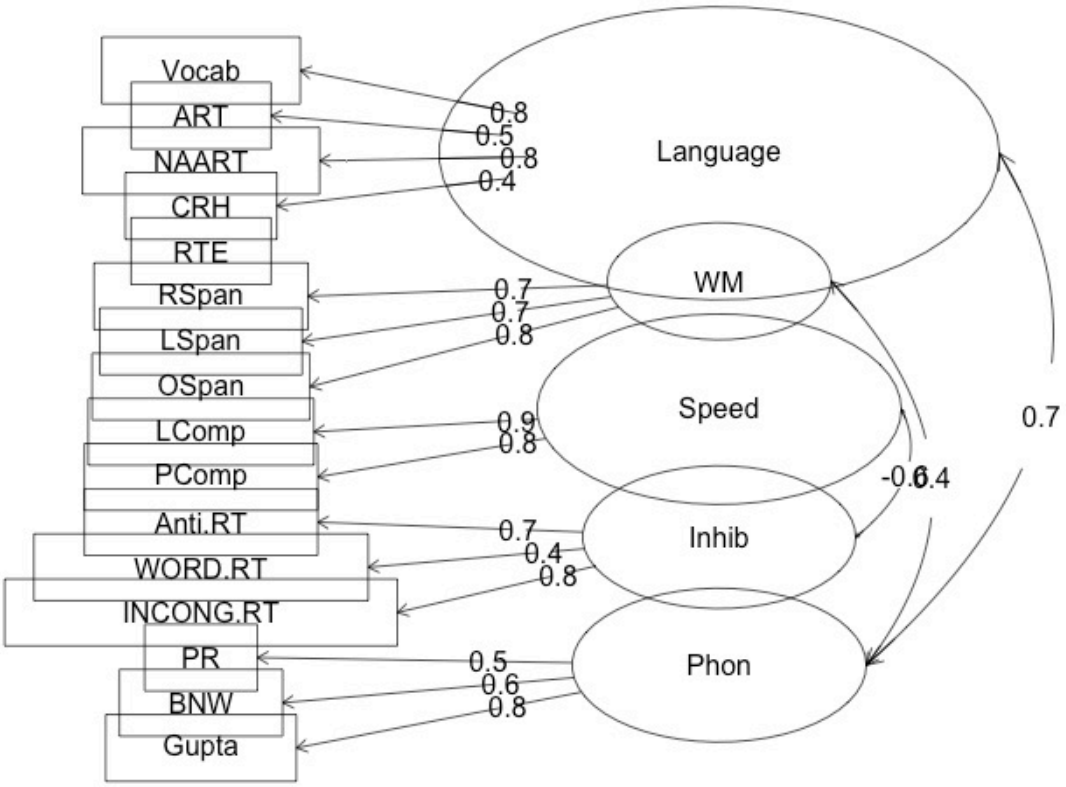


Figure 7. Estimated confirmatory factor analysis (CFA) five-factor model structure for predictor variables.

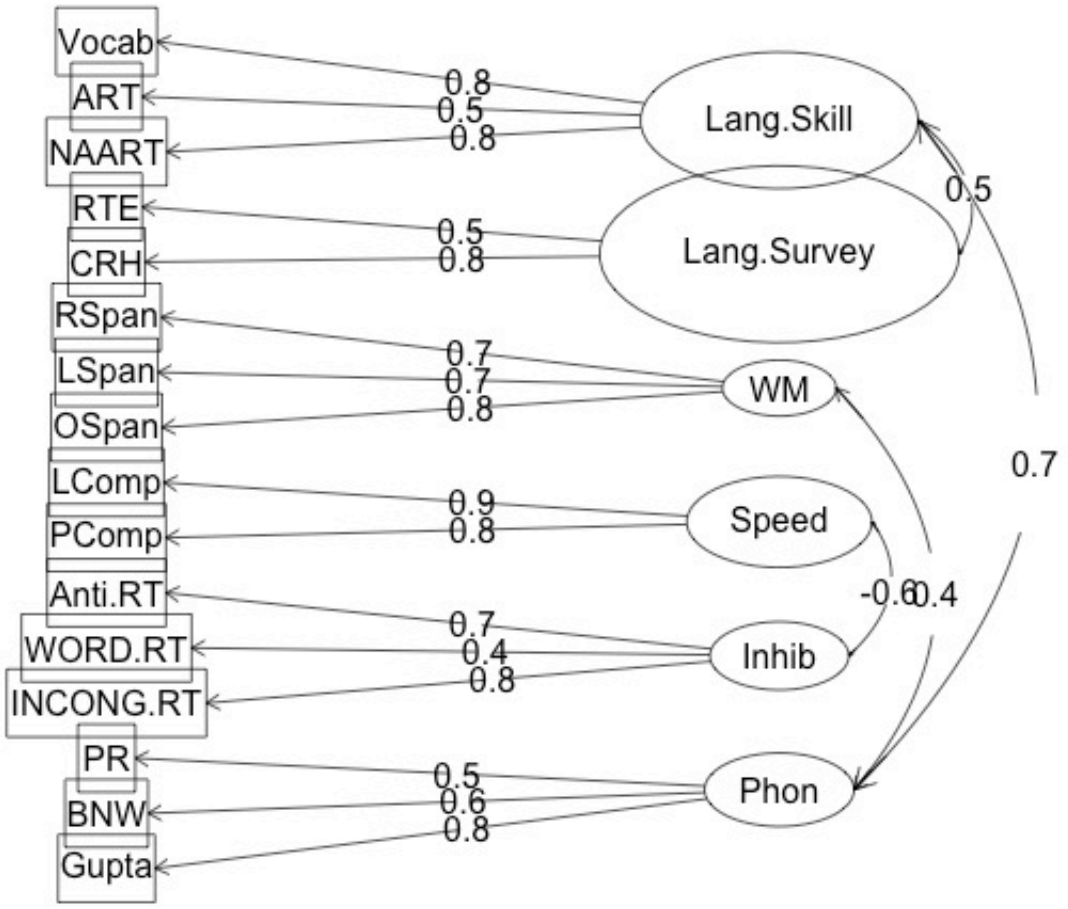


Figure 8. Estimated confirmatory factor analysis (CFA) six-factor model structure for predictor variables.

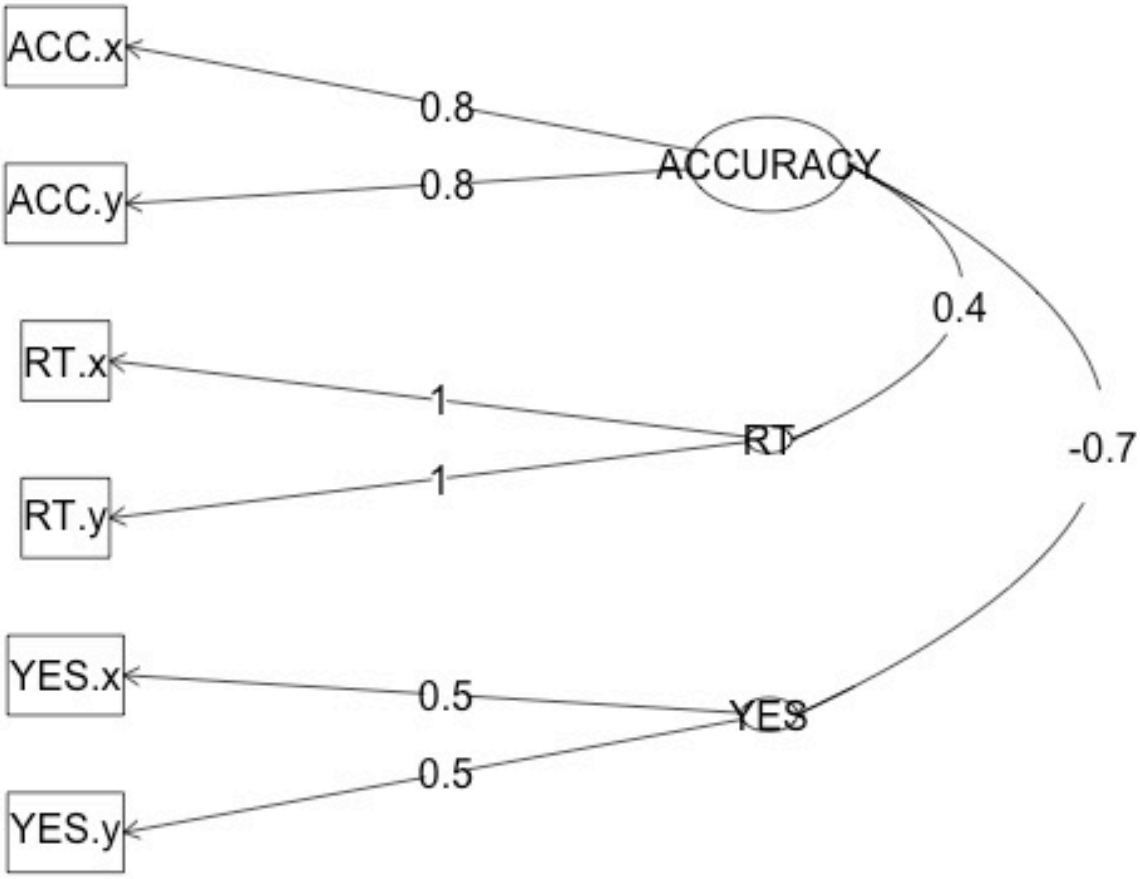


Figure 9. Estimated confirmatory factor analysis (CFA) model structure for overall reading outcomes.

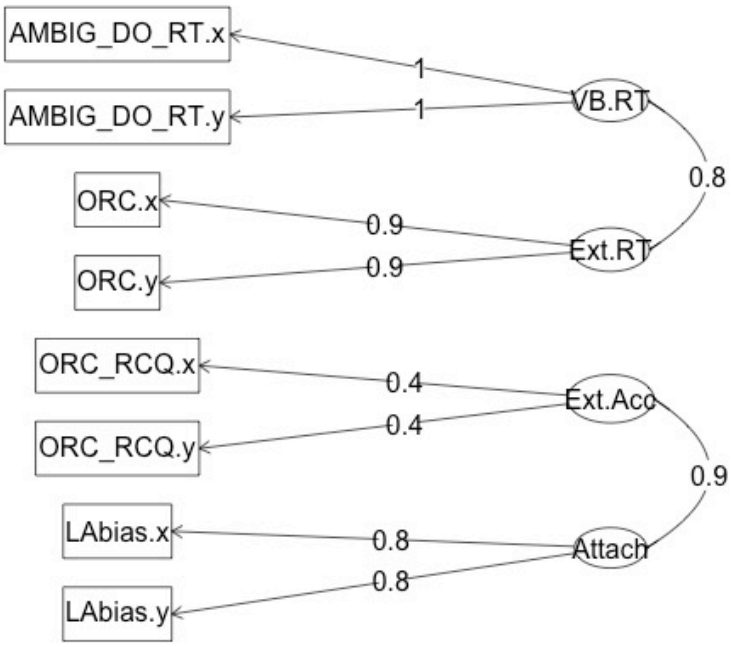


Figure 10. Estimated confirmatory factor analysis (CFA) model structure for syntactic outcome measures.

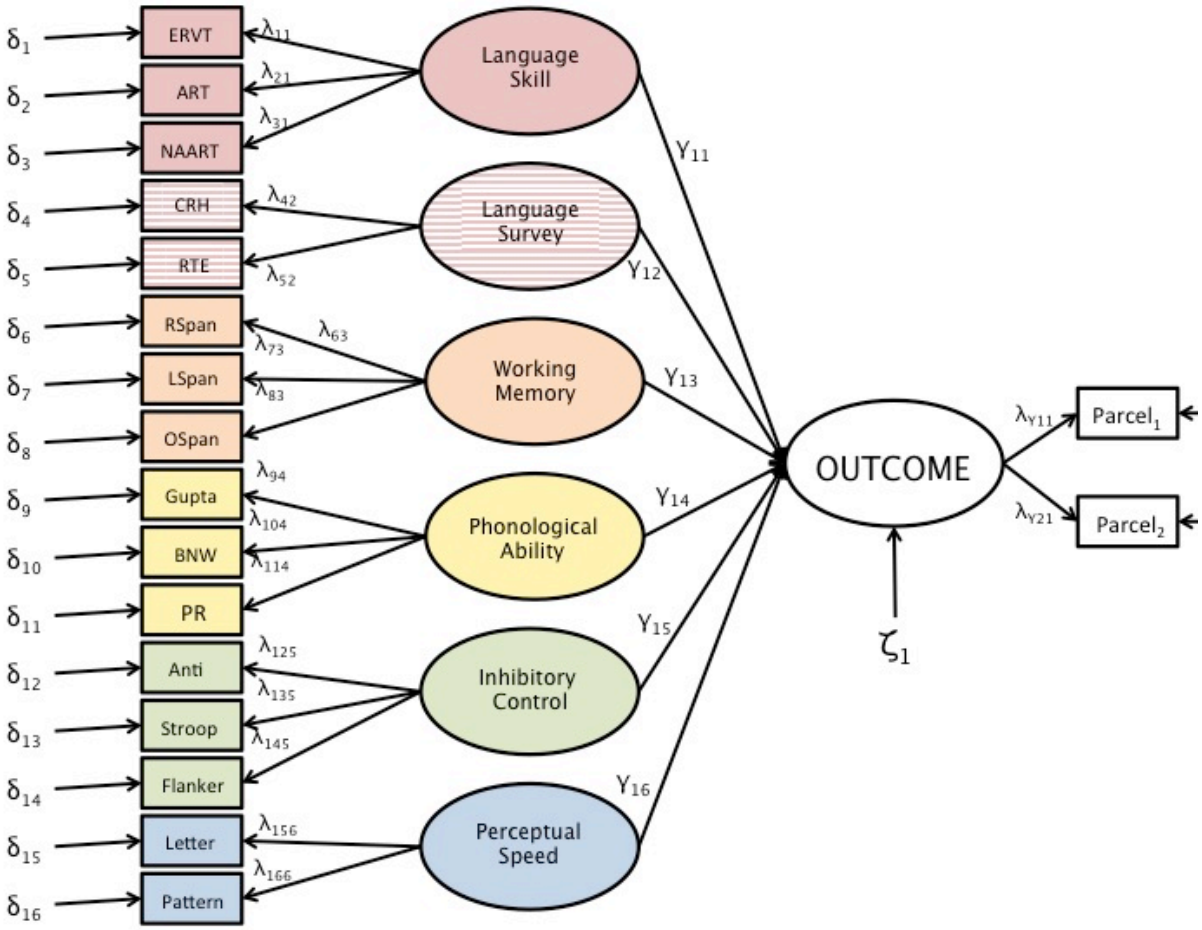


Figure 11. General structural equation model (SEM) diagram for critical analyses.

REFERENCES

Acheson, D. J., & MacDonald, M. C. (2011). The rhymes that the reader perused confused the meaning: Phonological effects during on-line sentence comprehension. *Journal of Memory and Language*, *65*, 193-207. doi: 10.1016/j.jml.2011.04.006

Acheson, D. J., Wells, J. B., & MacDonald, M. C. (2008). New and updated tests of print exposure and reading abilities in college students. *Behavior Research Methods*, *40*, 278–89. doi:10.3758/BRM.40.1.278

Altmann, G., & Steedman, M. (1988). Interaction with context during human sentence processing. *Cognition*, *30*, 191-238.

Arai, M., van Gompel, R. P. G., & Scheepers, C. (2007). Priming ditransitive structures in comprehension. *Cognitive Psychology*, *54*, 218-250. doi: 10.1016/j.cogpsych.2006.07.001

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

Babyak, M. A. (2004). What you see may not be what you get: a brief, nontechnical introduction to overfitting in regression-type models. *Psychosomatic Medicine*, *66*, 411-421.

Baddeley, A., Eldridge, M., & Lewis, V. (1981). The role of subvocalisation in reading. *The Quarterly Journal of Experimental Psychology*, *33*, 439-454.

Belsley, D. A., Kuh, E., & Welsch, R. E. (1980). *Regression diagnostics: Identifying influential data and sources of collinearity*. Hoboken, NJ: John Wiley & Sons.

Blair, J. R., & Spreen, O. (1989). Predicting premorbid IQ: a revision of the National Adult Reading Test. *The Clinical Neuropsychologist*, *3*(2), 129-136.

Bollen, K. A. (1989). *Structural equations with latent variables*. New York: John Wiley & Sons.

Borovsky, A., Elman, J. L., & Fernald, A. (2012). Knowing a lot for one's age: Vocabulary skill and not age is associated with anticipatory incremental sentence interpretation in children and adults. *Journal of Experimental Child Psychology*, *112*, 417-436.

Brainard, D. H. (1997) The Psychophysics Toolbox, *Spatial Vision, 10,* 433-436. doi: 10.1163/156856897X00357

Brown-Schmidt, S. (2009). The role of executive function in perspective taking during online language comprehension. *Psychonomic Bulletin & Review*, *16*, 893-900. doi: 10.3758/PBR.16.5.893

Brown-Schmidt, S., & Fraundorf, S. H. (2015). Interpretation of informational questions modulated by joint knowledge and intonational contours. *Journal of Memory and Language*, *84*, 49-74. doi: 10.1016/j.jml.2015.05.002

Byrne, B., & Letz, J. (1983). Phonological awareness in reading disabled adults. Australian Journal of Psychology, 35, 185-197.

Caplan, D., DeDe, G., Waters, G., Michaud, J., & Tripodis, Y. (2011). Effects of age, speed of processing, and working memory on comprehension of sentences with relative clauses. *Psychology and Aging*, *26*, 439.

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

Carter, B.T., & Luke, S.G. (2016). Individual differences in eye movements are consistent across time in reading. Poster presented at the 57th Annual Meeting of the Psychonomic Society, Boston, MA.

Christianson, K., Hollingworth, A., Halliwell, J. F., & Ferreira, F. (2001). Thematic roles assigned along the garden path linger. *Cognitive Psychology*, *42*, 368-407. doi: 10.1006/cogp.2001.0752

Coffman, D. L., & MacCallum, R. C. (2005). Using parcels to convert path analysis models into latent variable models. *Multivariate Behavioral Research*, *40*(2), 235-259.

Cohen, J. (1983). The cost of dichotomization. *Applied Psychological Measurement*, *7*, 249-253.

Cole, D. A., & Preacher, K. J. (2014). Manifest variable path analysis: Potentially serious and misleading consequences due to uncorrected measurement error. *Psychological Methods*, *19*(2), 300.

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

Coté, N., Goldman, S. R., & Saul, E. U. (1998). Students making sense of informational text: Relations between processing and representation. *Discourse Processes*, *25*, 1-53.

Cronbach, L. J. (1957). The two disciplines of scientific psychology. *American Psychologist*, *12*, 671-684.

Cuetos, F., & Mitchell, D. C. (1988). Cross-linguistic differences in parsing: Restrictions on the use of the Late Closure strategy in Spanish. *Cognition*, *30*, 73-105. doi: 10.1016/0010-0277(88)90004-2

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

Daneman, M., & Merikle, P. M. (1996). Working memory and language comprehension: A meta-analysis. *Psychonomic Bulletin & Review*, *3*, 422-433.

Duckworth, A. L., & Kern, M. L. (2011). A meta-analysis of the convergent validity of self-control measures. *Journal of Research in Personality*, *45*, 259-268.

Eriksen, C. W. (1995). The flankers task and response competition: A useful tool for investigating a variety of cognitive problems. *Visual Cognition*, *2*, 101-118.

Eriksen, B. A., & Eriksen, C. W. (1974). Effects of noise letters upon the identification of a target letter in a nonsearch task. *Perception & Psychophysics*, *16*, 143-149.

Ekstrom, R. B., French, J. W., Harman, H. H., & Dermen, D. (1976). *Kit of Factor-Referenced Cognitive Tests*. Princeton, NJ: Educational TestingSservice.

Engle, R. W. (2002). Working memory capacity as executive attention. *Current Directions in Psychological Science*, *11*, 19-23. doi: 10.1111/1467-8721.00160

Engle, R. W., Nations, J. K., & Cantor. J. (1990). Is "working memory capacity" just another name for word knowledge? *Journal of Educational Psychology*, *82*, 799-804.

Farmer, T. A., Fine, A. B., Misyak, J. B., & Christiansen, M. H. (2016). Reading span task performance, linguistic experience, and the processing of unexpected syntactic events. *The Quarterly Journal of Experimental Psychology*, *218*, 1–21. doi: 10.1080/17470218.2015.1131310

Farmer, T.A., Fine, A.B., Yan, S., Cheimariou, S., & Jaeger, T.F. (2014). Syntactic expectation adaptation in the eye-movement record. In P. Bello, M. Guarini, M. McShane, & B. Scassellati (Eds.), *Proceedings of the 36th Annual Meeting of the Cognitive Science Society* (pp. 2181-2186). Austin, TX: Cognitive Science Society.

Fedorenko, E., Gibson, E., & Rohde, D. (2006). The nature of working memory in sentence comprehension: Evidence against domain-specific working memory resources. *Journal of Memory and Language*, *54*, 541-553.

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*, 246-269, doi: 10.1016/j.jml.2006.06.007

Ferreira, F., & Clifton, C., Jr. (1986). The independence of syntactic processing. *Journal of Memory and Language*, *25*, 348-368. doi: 10.1016/0749-596X(86)9006-9

Fine, A.B., Jaeger, T.F., Farmer, T.A., & Qian, T. (2013). Rapid expectation adaptation during syntactic comprehension. *PLoS ONE*, *8*, e77661. doi: 10.1371/ journal.pone.0077661

Fine, A. B., Qian, T., Jaeger, T. F., & Jacobs, R. A. (2010). Is there syntactic adaptation in language comprehension? In *Proceedings of the 2010 Workshop on Cognitive Modeling and Computational Linguistics* (pp. 18-26). Uppsala, Sweden: Association for Computational Linguistics.

Fodor, J. A. (1983). *The modularity of mind: An essay on faculty psychology*. MIT press.

Fraundorf, S. H., Watson, D. G., & Benjamin, A. S. (2010). Recognition memory reveals just how CONTRASTIVE contrastive accenting really is. *Journal of Memory and Lang*uage, 63, 367-386. doi: 10.1016/jml.2010.06.004.

Fraundorf, S. H., Benjamin, A. S., & Watson, D. G. (2013). What happened (and what did not): Discourse constraints on encoding of plausible alternatives. *Journal of Memory and Language*, *69*, 196-227. doi: 10.1016/j.jml.2013.06.003

Fraundorf, S. H., & Jaeger, T. F. (2016). Readers generalize adaptation to newly-encountered dialectal structures to other unfamiliar structures. *Journal of Memory and Language, 91*, 28-58. doi: 10.1016/j.jml.2016.05.006

Fraundorf, S. H., & Watson, D. G. (2013). Alice's adventures in *um*-derland: Psycholinguistic sources of variation in disfluency production. *Language and Cognitive Processes*, *29*, 1083-1096. doi: 10.1080/01690965.2013.832785

Frazier, L., Carlson, K., & Clifton, C. (2006). Prosodic phrasing is central to language comprehension. *Trends in Cognitive Sciences*, *10*, 244-249.

Friedman, N. P., & Miyake, A. (2004). The reading span task and its predictive power for reading comprehension ability. *Journal of Memory and Language*, *51*, 136-158. doi: 10.1016/j.jml.2004.03.008

Garnsey, S. M., Pearlmutter, N. J., Myers, E., & Lotocky, M. A. (1997). The contributions of verb bias and plausibility to the comprehension of temporarily ambiguous sentences. *Journal of Memory and Language*, *37*, 58-93.

Gernsbacher, M. A. (1993). Less skilled readers have less efficient suppression mechanisms. *Psychological Science*, *4*, 294-298.

Gibson, E. (1998). Linguistic complexity: Locality of syntactic dependencies. *Cognition*, *68*, 1-76. doi: 10.1016/S0010-0277(98)00034-1

Gibson, E. (2000). The dependency locality theory: A distance-based theory of linguistic complexty. In A. Marantz, Y. Miyashita, & W. O'Neil (Eds.), *Image, language, brain: Papers from the first mind articulation project symposium* (pp. 94-126). Cambridge, MA: The MIT Press.

Gibson, E., Desmet, T., Grodner, D., Watson, D., & Ko, K. (2005). Reading relative clauses in English. *Cognitive Linguistics, 16*, 313-354.

Green, D. M., & Swets, J. A. (1966). Signal detection theory and psychophysics. New York: Wiley.

Gordon, P. C., Hendrick, R., & Johnson, M. (2001). Memory interference during language processing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *27*, 1-13.

Gupta, P. (2003). Examining the relationship between word learning, nonword repetition, and immediate serial recall in adults. *The Quarterly Journal of Experimental Psychology*, *56A*, 1213-1236. doi: 10.1080/02724980343000071

Hausmann, R. G., Vuong, A., Towle, B., Fraundorf, S. H., Murray, R. C., & Connelly, J. (2013, July). An evaluation of the effectiveness of just-in-time hints. In *International Conference on Artificial Intelligence in Education* (pp. 791-794). Springer Berlin Heidelberg.

Huettig, F., & Janse, E. (2016). Individual differences in working memory and processing speed predict anticipatory spoken language processing in the visual world. *Language, Cognition and Neuroscience*, *31*, 80–93. http://doi.org/10.1080/23273798.2015.1047459

Jarrold, C., & Towse, J. N. (2006). Individual differences in working memory. *Neuroscience*, *139*, 39-50.

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

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

Kane, M. J., Bleckley, M. K., Conway, A. R. A., & Engle, R. W. (2001). A controlled-attention view of working-memory capacity. *Journal of Experimental Psychology: General*, *130*, 169-183. doi: 10.1037//0096-3445.130.2.169

Kane, M. J., Conway, A. R., Hambrick, D. Z., & Engle, R. W. (2007). Variation in working memory capacity as variation in executive attention and control. In A.R. Conway, C.

Jarrold, M.J. Kane, A. Miyake, & J.N. Towse (Eds.), *Variation in working memory* (pp. 21-48). New York: Oxford University Press.

Kane, M. J., Hambrick, D. Z., Tuholski, S. W., Wilhelm, O., Payne, T. W., & Engle, R. W. (2004). The generality of working memory capacity: A latent-variable approach to verbal and visuospatial memory span and reasoning. *Journal of Experimental Psychology: General*, *133*, 189-217. doi: 10.1037/0096-3445.133.2.189

Kaschak, M.P. (2006). What this construction needs is generalized. *Memory & Cognition*, *34*, 368-379.

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

Keller, T. A., Carpenter, P. A., & Just, M. A. (2003). Brain imaging of tongue-twister sentence comprehension: Twisting the tongue and the brain. *Brain and Language*, *84*, 189-203.

Kennison, S. M. (2001). Limitations on the use of verb information during sentence comprehension. *Psychonomic Bulletin & Review*, *8*, 132-138.

Kennison, S. M. (2004). The effect of phonemic repetition on syntactic ambiguity resolution: Implications for models of working memory. *Journal of Psycholinguistic Research*, *33*, 493-516.

Kenny, D. A. (1979). *Correlation and causation*. New York: John Wiley.

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

Kleiner, M., Brainard, D., & Pelli, D. (2007). What's new in Psychtoolbox-3? *Perception 36* ECVP Abstract Supplement.

Kline, R. B. (2016). *Principles and practice of structural equation modeling* (Fourth edition). New York: The Guildford Press.

Kuperman, V., & Van Dyke, J. A. (2011). Effects of individual differences in verbal skills on eye-movement patterns during sentence reading. *Journal of Memory and Language*, *68*, 42-73. doi: 10.1016/j.jml.2011.03.002

Kush, D., Johns, C. L., & Van Dyke, J. A. (2015). Identifying the role of phonology in sentence-level reading. *Journal of Memory and Language*, *79*, 18-29.

Lau, E. F., & Ferreira, F. (2005). Lingering effects of disfluent material on comprehension of garden path sentences. *Language and Cognitive Processes*, *20*, 633-666. doi: 10.1080/01690960444000142

Lee, E.-K., & Fraundorf, S. H. (in press). Effects of contrastive accents in memory for L2 discourse. *Bilingualism: Language and Cognition*.

Lee, E.-K., Lu, D. H.-Y., & Garnsey, S. M. (2013). L1 word order and sensitivity to verb bias in L2 processing. *Bilingualism: Language and Cognition*, *16*, 761-775.

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, *106*, 1126-1177. doi: 10.1016/j.cognition.2007.05.006

Levy, R., Fedorenko, E., Breen, M., & Gibson, E. (2012). The processing of extraposed structures in English. *Cognition*, *122*, 12-36. doi: 10.1016/j.cognition.2011.07.012

Lewis, R. L., Vasishth, S., & Van Dyke, J. A. (2006). Computational principles of working memory in sentence comprehension. *TRENDS in Cognitive Sciences,* 10, 447-454. doi: 10.1016/j.tics.2006.08.007

Long, D. L., & Prat, C. S. (2008). Individual differences in syntactic ambiguity resolution: Readers vary in their use of plausibility information. *Memory & Cognition*, *36*, 375-391.

Lord, F. M. (1963). Elementary models for measuring change. In C. W. Harris (Ed.), *Problems in measuring change* (pp. 21–38). Madison, WI: University of Wisconsin Press.

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

Luka, B. J., & Choi, H. (2012). Dynamic grammar in adults: Incidental learning of natural syntactic structures extends over 48 h. *Journal of Memory and Language*, *66*, 345-360.

Lustig, C., May, C. P., & Hasher, L. (2001). Working memory span and the role of proactive interference. *Journal of Experimental Psychology: General*, *130*, 199-207. doi: 10.1037//0096-3445.130.2.199

MacCallum, R. C., Zhang, S., Preacher, K. J., & Rucker, D. D. (2002). On the practice of dichotomization of quantitative variables. *Psychological Methods*, *7*, 19-40.

MacDonald, M. C. (1994). Probabilistic constraints and syntactic ambiguity resolution. *Language and Cognitive Processes*, *9*, 157-201.

MacDonald, M. C., Just, M. A., & Carpenter, P. A. (1992). Working memory constraints on the processing of syntactic ambiguity. *Cognitive Psychology*, *24*, 56-98.

, M. C., & Christiansen, M. H. (2002). Reassessing working memory: Comments on Just and Carpenter (1992) and Waters and Caplan (1996). *Psychological Review*, *109*, 35-54. doi:10.1037//033-295X.109.1.35

MacDonald, M. C., Pearlmutter, N. J., & Seidenberg, M. S. (1994). The lexical nature of syntactic ambiguity resolution. *Psychological Review*, *101*, 676-703.

Maclay, H., & Osgood, C. E. (1959). Hesitation phenomena in spontaneous speech. *Word*, *14*, 19-44.

Macmillan, N. A., & Creelman, C. D. (2005). *Detection theory* (2nd ed.). New York: Erlbaum.

Martin, A.E., & McElree, B. (2008). A content-addressable pointer mechanism underlies comprehension of verb-phrase ellipsis. *Journal of Memory and Language*, *58*, 879-906. doi: 10.1016/j.jml.2007.06.010

Matsuki, K., Kuperman, V., & Van Dyke, J. A. (2016). The random forests statistical technique: An examination of its value for the study of reading. *Scientific Studies of Reading*, 20, 20-33. doi: 10.1080/10888438.2015.1107073.

McCabe, D. (2010). The influence of complex working memory span task administration methods on prediction of higher level cognition and metacognitive control of response times. *Memory & Cognition*, *38*, 868-882. doi: 10.3758/MC.38.7.868

McCutchen, D., Bell, L. C. France, I. M., & Perfetti, C.A (1991). Phoneme-specific interference in reading: The tongue-twister effect revisited. *Reading Research Quarterly*, *26*, 87-103.

McElree, B., Foraker, S., & Dyer, L. (2003). Memory structures that subserve sentence comprehension. *Journal of Memory and Language*, *48*, 67-91. doi: 10.1016/S0749-596X(02)00515-6

McNamara, D. S., Kintsch, E., Songer, N. B., & Kintsch, W. (1996). Are good texts always better? Interactions of text coherence, background knowledge, and levels of understanding in learning from text. *Cognition and Instruction*, *14*, 1-43.

Mishra, R. K., Singh, N., Pandey, A., & Huettig, F. (2012). Spoken language-mediated anticipatory eye-movements are modulated by reading ability-Evidence from Indian low and high literates. *Journal of Eye Movement Research*, *5*. doi:10.16910/jemr.5.1.3

Miyake, A., Friedman, N. P., Emerson, M. J., Witzki, A. H., Howerter, A., & Wager, T. D. (2000). The unity and diversity of executive functions and their contributions to complex "frontal lobe" tasks: A latent variable analysis. *Cognitive psychology*, *41*(1), 49-100.

Mueller, R. O., & Hancock, G. R. (2008). Best practices in structural equation modeling. *Best practices in quantitative methods*, 488-508.

Murayama, K., Sakaki, M., Yan, V. X., & Smith, G. M. (2014). Type I error inflation in the traditional by-participant analysis to metamemory accuracy: A generalized mixed-effects model perspective. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *40*, 1287-1306.

Nairne, J. S. (2002). Remembering over the short-term: The case against the standard model. *Annual Review of Psychology*, *53*, 53-81. doi: 10.1146/annurev.psych.53.100901.135131

Novick, J. M., Trueswell, J. C., & Thompson-Schill, S. L. (2005). Cognitive control and parsing: Reexamining the role of Broca's area in sentence comprehension. Cognition, Affective, & Behavioral Neuroscience, 5, 263-281. doi: 10.3758/CABN.5.3.263.

Novick, J. M., Trueswell, J. C., & Thompson-Schill, S. L. (2010). Broca's area and language processing: Evidence for the cognitive control connection. *Language and Linguistics Compass*, *4*, 906-924. doi: 10.1111/j.1749-818x.2010.00244.

Pashler, H., McDaniel, M., Rohrer, D., & Bjork, R. (2008). Learning styles: Concepts and evidence. *Psychological Science in the Public Interest*, *9*, 105-119.

Payne, B.R., Grison, S., Gao, X., Christianson, K., Morrow, D.G., & Stine-Morrow, E.A.L. (2014). Aging and individual differences in binding during sentence understanding: Evidence from temporary and global syntactic attachment ambiguities. *Cognition*, *130*, 157-173.

Pearlmutter, N. J., & MacDonald, M. C. (1995). Individual differences and probabilistic constraints in syntactic ambiguity resolution. *Journal of Memory and Language*, *34*, 521-542.

Pellegrino, J. W., Baxter, G. P., & Glaser, R. (1999). Addressing the" two disciplines" problem: Linking theories of cognition and learning with assessment and instructional practice. *Review of Research in Education*, *24*, 307-353.

Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision, 10*, 437-442. doi: 10.1163/156856897X00357

Perfetti, C. A., & Hart, L. (2002). The lexical quality hypothesis. In L. Verhoeven, C. Elbro, & P. Reitsma (Eds.), *Precursors of functional literacy* (pp. 67-86). Philadelphia, PA: John Benjamins Publishing Company.

Preacher, K. J., Rucker, D. D., MacCallum, R. C., & Nicewander, W. A. (2005). Use of the extreme groups approach: a critical reexamination and new recommendations. *Psychological Methods*, *10*, 178-192.

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, *124*, 372-422.

Rayner, K., Carlson, M., & Frazier, L. (1983). The interaction of syntax and semantics during sentence processing: Eye movements in the analysis of semantically biased sentences. *Journal of Verbal Learning and Verbal Behavior*, *22*, 358-374.

Read, C., & Ruyter, L. (1985). Reading and spelling skills in adults of low literacy. *Remedial and Special Education*, *6*, 43-52.

Redick, T.S., & Engle, R.W. (2006). Working memory capacity and attention network test performance. *Applied Cognitive Psychology*, *20*, 713-721. doi: 10.1002/acp.1224

Rommers, J., Meyer, A. S., & Huettig, F. (2015). Verbal and nonverbal predictors of language-mediated anticipatory eye movements. *Attention, Perception, & Psychophysics*, *77*, 720–730. http://doi.org/10.3758/s13414-015-0873-x

Ryskin, R. A., Benjamin, A. S., Tullis, J., & Brown-Schmidt, S. (2015). Perspective-taking in comprehension, production, and memory: An individual differences analysis. *Journal of Experimental Psychology: General*, *144*, 898-915. doi: 10.1037/xge0000093

Ryskin, R. A., Qi, Z., Duff, M. C., & Brown-Schmidt, S. (2016). Verb biases are shaped through lifelong learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition.* doi: 10.1037/xlm0000341

Salthouse, T. A. (1996). The processing-speed theory of adult age differences in cognition. *Psychological Review*, *103*, 403-428.

Salthouse, T. A., & Babcock, R. L. (1991). Decomposing adult age differences in working memory. *Developmental Psychology*, *27*, 763.

Salthouse, T. A., & Pink, J. E. (2008). Why is working memory related to fluid intelligence? *Psychonomic Bulletin & Review*, *15*, 364-371. doi: 10.3758/PBR.15.2.364

Salthouse, T. A., Siedlecki, K. L., & Krueger, L. E. (2006). An individual differences analysis of memory control. *Journal of Memory and Language*, *55*, 102-125. doi: 10.1016/j.jml.2006.03.006

Sawyer, D. J., & Fox, B. J. (1991). *Phonological awareness in reading: The evolution of current perspectives*. Springer series in language and communication (Vol. 28). New York: Springer-Verlag, Inc.

Seidenberg, M. S. (1985). The time course of phonological code activation in two writing systems. *Cognition*, *19*, 1-30.

Shah, P., & Miyake, A. (1996). The separability of working memory resources for spatial thinking and language processing: an individual differences approach. *Journal of Experimental Psychology: General*, *125*, 4-27.

Spearman, C. (1904). The proof and measurement of association between two things. *The American Journal of Psychology*, *15*, 72-101.

Spivey-Knowlton, M. J., Trueswell, J. C., & Tanenhaus, M. K. (1993). Context effects in syntactic ambiguity resolution: Discourse and semantic influences in parsing reduced relative clauses. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, *47*, 276-309.

Stanovich, K. E. (1985). Explaining the variance in reading ability in terms of psychological processes: What have we learned?. *Annals of Dyslexia*, *35*, 67-96.

Stanovich, K. E., & West, R. F. (1989). Exposure to print and orthographic processing. *Reading Research Quarterly*, 24, 402-433.

Stine, E. A. L., & Hindman, J. (1994). Age differences in reading time allocation for propositionally dense sentences. *Aging and Cognition*, *1*, 2-16. doi:10.1080/09289919408251446

Stine-Morrow, E. A. L., Soederberg Miller, L. M., Gagne, D. D., & Hertzog, C. (2008). Self-regulated reading in adulthood. *Psychology and Aging*, *23*, 131-153. doi: 10.1037/0882-7974.23.1.131

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology*, *18*, 643-662.

Swets, B., Desmet, T., Hambrick, D. Z., & Ferreira, F. (2007). The role of working memory in syntactic ambiguity resolution: A psychometric approach. *Journal of Experimental Psychology: General*, *136*, 64-81. doi: 10.1037/0096-3445.136.1.64

Tabor, W., Galantucci, B., & Richardson, D. (2004). Effects of merely local syntactic coherence on sentence processing. *Journal of Memory and Language*, *50*, 355-370. doi: 10.1016/j.jml.2004.01.001

Tanner, D., & Bulkes, N. Z. (2015). Cues, quantification, and agreement in language comprehension. *Psychonomic Bulletin & Review*, *22*, 1753-1763.

Thompson, B. (2000). Ten commandments of structural equation modeling. In L. G. Grimm & P. R. Yarnold (Eds.), Reading and understanding more multivariate statistics (pp. 261-283). Washington, DC: American Psychological Association.

Thothathiri, M., & Snedeker, J. (2008). Give and take: Syntactic priming during spoken language comprehension. *Cognition*, *108*, 51-68. doi: 10.1016/j.cognition.2007.12.012

Tokowicz, N., & MacWhinney, B. (2005). Implicit and explicit measures of sensitivity to violations in second language grammar: An event-related potential investigation. *Studies in Second Language Acquisition*, *27*, 173-204. doi: 10.1017/S0272263105050102

Tooley, K. M., & Traxler, M. J., & Swaab, T. Y. (2009). Electrophysiological and behavioral evidence of syntactic priming in sentence comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*, 19-45. doi: 10.1037/a0013984

Traxler, M. J. (2008). Lexically independent priming in online sentence comprehension. *Psychonomic Bulletin & Review*, 15, 149-155. doi: 10.3758/PBR.15.1.149.

Traxler, M. J. (2007). Working memory contributions to relative clause attachment processing: A hierarchical linear modeling analysis. *Memory & Cognition*, 35, 1107-1121. doi: 10.3758/BF03193482.

Traxler, M. J., & Tooley, K. M. (2007). Lexical mediation and context effects in sentence processing. *Brain Research*, *1146*, 59-74. doi: 10.1016/j.brainres.2006.10.010

Traxler, M. J., Williams, R. S., Blozis, S. A., & Morris, R. K. (2005). Working memory, animacy, and verb class in the processing of relative clauses. *Journal of Memory and Language, 53,* 204-224.

Turner, M. L., & Engle, R. W. (1989). Is working memory capacity task dependent? *Journal of Memory and Language*, *28*, 127-154.

United Nations Development Programme. (2011). Human development index. Retrieved from http://hdr.undp.org/en/statistics/hdi/

Unsworth, N., & Engle, R. W. (2007). The nature of individual differences in working memory: Active maintenance in primary memory and controlled search from secondary memory. *Psychological Review*, *114*, 104-132. doi: 10.1037/0033-295X.114.1.104

Unsworth, N., Heitz, R. P., Schrock, J. C., & Engle, R. W. (2005). An automated version of the operation span task. *Behavior Research Methods*, *37*, 498-505.

Unsworth, N., Redick, T. S., Heitz, R. P., Broadway, J. M., & Engle, R. W. (2009). Complex working memory span tasks and higher-order cognition: A latent-variable analysis of the relationship between processing and storage. *Memory, 17*, 635-654. doi: 10.1080/09658210902998047

Uttl, B. (2002). North American Adult Reading Test: age norms, reliability, and validity. *Journal of Clinical and Experimental Neuropsychology*, *24*(8), 1123-1137.

Van Dyke, J. A., Johns, C. L., & Kukona, A. (2014). Low working memory capacity is only spuriously related to poor reading comprehension. *Cognition*, *131*, 373-403.

Van Zandt, T. (2000). How to fit a response time distribution. *Psychonomic Bulletin & Review*, *7*, 424-465.

Vogel, E. K., & Awh, E. (2008). How to exploit diversity for scientific gain: Using individual differences to constrain cognitive theories. *Current Directions in Psychological Science*, *17*, 171-176.

Wagner, R. K., & Torgesen, J. K. (1987). The nature of phonological awareness and its causal role in the acquisition of reading skills. *Psychological Bulletin*, *101*, 192–212.

Wagner, R., Torgesen, J., & Rashotte, C. (1999). *Comprehensive test of phonological processing (CTOPP)*. Austin, TX: Pro-Ed.

Waters, G. S., & Caplan, D. (1996). The capacity theory of sentence comprehension: Critique of Just and Carpenter (1992). *Psychological Review*, *103*, 761-772.

Waters, G. S., & Caplan, D. (2003). The reliability and stability of verbal working memory measures. *Behavior Research Methods, Instruments, & Computers*, *35*, 550-564.

Waters, G. S., & Caplan, D. (2005). The relationship between age, processing speed, working memory capacity, and language comprehension. *Memory*, *13*, 403-413. doi: 10.1080/09658210344000459

Wells, J. B., Christiansen, M. H., Race, D. S., Acheson, D. J., & MacDonald, M. C. (2009). Experience and sentence processing: Statistical learning and relative clause comprehension. *Cognitive Psychology*, *58*, 250-271. doi: 10.1016/j.cogpsych.2008.08.002

Westfall, J., & Yarkoni, T. (2016). Statistically controlling for confounding constructs is harder than you think. *PLoS ONE*, *11*, e0152719.

Williams, L. J., & O'Boyle, E. H. (2008). Measurement models for linking latent variables and indicators: A review of human resource management research using parcels. *Human Resource Management Review*, *18*(4), 233-242.

Wilson, M. P., & Garnsey, S. M. (2009). Making simple sentences hard: Verb bias effects in simple direct object sentences. *Journal of Memory and Language*, *60*, 368-392.

Wöstmann, N. M., Aichert, D. S., Costa, A., Rubia, K., Möller, H. J., & Ettinger, U. (2013). Reliability and plasticity of response inhibition and interference control. *Brain and Cognition*, *81*, 82-94.

APPENDIX A: MIXED-EFFECT MODEL EQUATIONS FOR MODELS OF SYNTACTIC EFFECTS WITH EXPERIMENTAL CONDITIONS ONLY

Residual reading time for verb bias items was modeled as:

$$Y_{ij} = \gamma_{000} + \gamma_{100}*\text{Ambiguity} + \gamma_{200}*\text{Bias} + \gamma_{1200}*\text{Ambiguity*Bias} + u_{0i0} + u_{1i0}*\text{Ambiguity} + u_{2i0}*\text{Bias} + u_{12i0}*\text{Ambiguity*Bias} + v_{00j} + \varepsilon_{ij}$$

where Y_{ij} is the residual reading time for subject i on item j, γ_{000} is an intercept term representing grand mean residual reading time, γ_{100}, γ_{200}, and γ_{1200} are the fixed effects of the experimental conditions, u_{0i0} is the error for the subject intercept for subject i, u_{1i0}, $u_{2i0,}$ and u_{12i0} are the error terms for the random subject slopes for the experimental conditions for subject i, v_{00j} is the error for the item intercept for item j, and ε_{ij} is the error for subject i on item j.

Comprehension accuracy for verb bias items was modeled as:

$$\log(Y_{ij}) = \gamma_{000} + \gamma_{100}*\text{Ambiguity} + \gamma_{200}*\text{Bias} + \gamma_{1200}*\text{Ambiguity*Bias} + u_{0i0} + u_{1i0}*\text{Ambiguity} + u_{2i0}*\text{Bias} + u_{12i0}*\text{Ambiguity*Bias} + v_{00j} + v_{10j}*\text{Ambiguity}$$

where Y_{ij} are the odds of subject i correctly responding to item j, γ_{000} is an intercept term representing grand mean accuracy, γ_{100}, γ_{200}, and γ_{1200} are the fixed effects of the experimental conditions, u_{0i0} is the error for the subject intercept for subject i, u_{1i0}, $u_{2i0,}$ and u_{12i0} are the error terms for the random subject slopes for the experimental conditions for subject i, v_{00j} is the error for the item intercept for item j, and v_{10j} is the error for the random slope of ambiguity for item j, and ε_{ij} is the error for subject i on item j.

Residual reading time for relative-clause extraction items was modeled as:

$$Y_{ij} = \gamma_{000} + \gamma_{100}*\text{RCType} + u_{0i0} + u_{1i0}*\text{RCType} + v_{00j} + v_{10j}*\text{RCType} + \varepsilon_{ij}$$

where Y_{ij} is the residual reading time for subject i on item j, γ_{000} is an intercept term representing grand mean residual reading time, γ_{100} is the fixed effect of experimental condition, u_{0i0} is the error for the subject intercept for subject i, u_{1i0} is the error term for the random slope of condition for subject i, v_{00j} is

the error for the item intercept for item *j*, v_{10j} is the error for the random slope of condition for item *j*, and ε_{ij} is the error for subject *i* on item *j*.

Comprehension accuracy for relative-clause extraction items was modeled as:

$$\log(Y_{ij}) = \gamma_{000} + \gamma_{100}*\text{RCType} + \gamma_{200}*\text{QuestionType} + \gamma_{1200}*\text{RCType}*\text{QuestionType} + u_{0i0} + u_{1i0}*\text{RCType} + u_{2i0}*\text{QuestionType} + u_{12i0}*\text{RCType}*\text{QuestionType} + v_{00j} + v_{10j}*\text{RCType} + v_{20j}*\text{QuestionType} + v_{120j}*\text{RCType}*\text{QuestionType}$$

where Y_{ij} are the odds of subject *i* correctly responding to item *j*, γ_{000} is an intercept term representing grand mean accuracy, γ_{100}, γ_{200}, and γ_{1200} are the fixed effects of the experimental conditions, u_{0i0} is the error for the subject intercept for subject *i*, u_{1i0}, u_{2i0}, and u_{12i0} are the error terms for the random subject slopes for the experimental conditions for subject *i*, v_{00j} is the error for the item intercept for item *j*, and v_{10j}, v_{20j}, and v_{120j} are the error terms for the random slopes for the experimental conditions for item *j*, and ε_{ij} is the error for subject *i* on item *j*.

Responses to the attachment ambiguity items were modeled as:

$$\log(Y_{ij}) = \gamma_{000} + \gamma_{100}*\text{QuestionType} + u_{0i0} + u_{1i0}*\text{QuestionType} + v_{00j} + v_{10j}*\text{QuestionType}$$

where Y_{ij} are the odds of subject *i* answering *yes* to item *j*, γ_{000} is an intercept term representing the grand mean of answering *yes* (response bias), γ_{100} is the fixed effect of the question type condition (low- or high-attachment) on *yes* responses (sensitivity), u_{0i0} is the error for the subject intercept for subject *i*, u_{1i0} is the error term for the random subject slope of condition for subject *i*, v_{00j} is the error for the item intercept for item *j*, and v_{10j} is the error term for the random slope of condition for item *j*.

APPENDIX B: MIXED-EFFECT MODEL EQUATIONS FOR MODELS OF SYNTACTIC EFFECTS WITH EXPERIMENTAL CONDITIONS AND INDIVIDUAL DIFFERENCES

Residual reading time for verb bias items was modeled as:

$$Y_{ij} = \gamma_{000} + \gamma_{100}*\text{Ambiguity} + \gamma_{200}*\text{Bias} + \gamma_{1200}*\text{Ambiguity}*\text{Bias} + \gamma_{300}*\text{VWM} + \gamma_{400}*\text{Inhib} + \gamma_{500}*\text{Phon} + \gamma_{600}*\text{Speed} + \gamma_{700}*\text{Lang} + \gamma_{1300}*\text{VWM}*\text{Ambiguity} + \gamma_{1400}*\text{Inhib}*\text{Ambiguity} + \gamma_{1500}*\text{Phon}*\text{Ambiguity} + \gamma_{1600}*\text{Speed}*\text{Ambiguity} + \gamma_{1700}*\text{Lang}*\text{Ambiguity} + \gamma_{2300}*\text{VWM}*\text{Bias} + \gamma_{2400}*\text{Inhib}*\text{Bias} + \gamma_{1500}*\text{Phon}*\text{Bias} + \gamma_{2600}*\text{Speed}*\text{Ambiguity} + \gamma_{2700}*\text{Lang}*\text{Ambiguity} + \gamma_{12300}*\text{VWM}*\text{Ambiguity}*\text{Bias} + \gamma_{12400}*\text{Inhib}*\text{Ambiguity}*\text{Bias} + \gamma_{12500}*\text{Phon}*\text{Ambiguity}*\text{Bias} + \gamma_{12600}*\text{Speed}*\text{Ambiguity}*\text{Bias} + \gamma_{12700}*\text{Lang}*\text{Ambiguity}*\text{Bias} + u_{0i0} + u_{1i0}*\text{Ambiguity} + u_{2i0}*\text{Bias} + u_{12i0}*\text{Ambiguity}*\text{Bias} + v_{00j} + \varepsilon_{ij}$$

where Y_{ij} is the residual reading time for subject i on item j, γ_{000} is an intercept term representing grand mean residual reading time, γ_{100}, γ_{200}, and γ_{1200} are the fixed effects of the experimental conditions, γ_{300} through γ_{700} are the fixed effects of the individual-difference composites on overall reading time, γ_{1300} through γ_{1700} are the fixed effects of the individual-difference composites on the ambiguity effect, γ_{2300} through γ_{2700} are the fixed effects of the individual-difference composites on the verb bias effect, γ_{12300} through γ_{12700} are the fixed effects of the individual-difference composites on the ambiguity x verb bias interaction, u_{0i0} is the error for the subject intercept for subject i, u_{1i0}, $u_{2i0,}$ and u_{12i0} are the error terms for the random subject slopes for the experimental conditions for subject i, v_{00j} is the error for the item intercept for item j, and ε_{ij} is the error for subject i on item j.

Comprehension accuracy for verb bias items was modeled as:

$$\log(Y_{ij}) = \gamma_{000} + \gamma_{100}*\text{Ambiguity} + \gamma_{200}*\text{Bias} + \gamma_{1200}*\text{Ambiguity}*\text{Bias} + \gamma_{300}*\text{VWM} + \gamma_{400}*\text{Inhib} + \gamma_{500}*\text{Phon} + \gamma_{600}*\text{Speed} + \gamma_{700}*\text{Lang} + \gamma_{1300}*\text{VWM}*\text{Ambiguity} + \gamma_{1400}*\text{Inhib}*\text{Ambiguity} + \gamma_{1500}*\text{Phon}*\text{Ambiguity} + \gamma_{1600}*\text{Speed}*\text{Ambiguity} + \gamma_{1700}*\text{Lang}*\text{Ambiguity} + \gamma_{2300}*\text{VWM}*\text{Bias} + \gamma_{2400}*\text{Inhib}*\text{Bias} + \gamma_{1500}*\text{Phon}*\text{Bias} + \gamma_{2600}*\text{Speed}*\text{Ambiguity} + \gamma_{2700}*\text{Lang}*\text{Ambiguity} + \gamma_{12300}*\text{VWM}*\text{Ambiguity}*\text{Bias} + \gamma_{12400}*\text{Inhib}*\text{Ambiguity}*\text{Bias} + \gamma_{12500}*\text{Phon}*\text{Ambiguity}*\text{Bias} +$$

γ_{12600}*Speed*Ambiguity*Bias + γ_{12700}*Lang*Ambiguity*Bias + u_{0i0} + u_{1i0}*Ambiguity + u_{2i0}*Bias + u_{12i0}*Ambiguity*Bias + v_{00j} + v_{10j}*Ambiguity

where Y_{ij} are the odds of subject i correctly responding to item j, γ_{000} is an intercept term representing grand mean accuracy, γ_{100}, γ_{200}, and γ_{1200} are the fixed effects of the experimental conditions, γ_{300} through γ_{700} are the fixed effects of the individual-difference composites on overall reading time, γ_{1300} through γ_{1700} are the fixed effects of the individual-difference composites on the ambiguity effect, γ_{2300} through γ_{2700} are the fixed effects of the individual-difference composites on the verb bias effect, γ_{12300} through γ_{12700} are the fixed effects of the individual-difference composites on the ambiguity x verb bias interaction, u_{0i0} is the error for the subject intercept for subject i, u_{1i0}, $u_{2i0,}$ and u_{12i0} are the error terms for the random subject slopes for the experimental conditions for subject i, v_{00j} is the error for the item intercept for item j, and v_{10j} is the error for the random slope of ambiguity for item j, and ε_{ij} is the error for subject i on item j.

Residual reading time for relative-clause extraction items was modeled as:

$Y_{ij} = \gamma_{000} + \gamma_{100}$*RCType + γ_{200}*VWM + γ_{300}*Inhib + γ_{400} *Phon + γ_{500}*Speed + γ_{600}*Lang + γ_{1200}*VWM*RCType + γ_{1300}*Inhib*RCType + γ_{1400} *Phon*RCType + γ_{1500}*Speed*RCType + γ_{1600}*Lang*RCType + u_{0i0} + u_{1i0}*RCType + v_{00j} + v_{10j}*RCType + ε_{ij}

where Y_{ij} is the residual reading time for subject i on item j, γ_{000} is an intercept term representing grand mean residual reading time, γ_{100} is the fixed effect of experimental condition, γ_{200} through γ_{600} are the fixed effects of the individual-difference composites on overall reading time, γ_{1200} through γ_{1600} are the fixed effects of the individual-difference composites on the RC type effect,
u_{0i0} is the error for the subject intercept for subject i, u_{1i0} is the error term for the random slope of condition for subject i, v_{00j} is the error for the item intercept for item j, v_{10j} is the error for the random slope of condition for item j, and ε_{ij} is the error for subject i on item j.

Comprehension accuracy for relative-clause extraction items was modeled as:

$$\log(Y_{ij}) = \gamma_{000} + \gamma_{100}*\text{RCType} + \gamma_{200}*\text{QuestionType} + \gamma_{1200}*\text{RCType}*\text{QuestionType} + u_{0i0} + u_{1i0}*\text{RCType} + u_{2i0}*\text{QuestionType} + u_{12i0}*\text{RCType}*\text{QuestionType} + \gamma_{300}*\text{VWM} + \gamma_{400}*\text{Inhib} + \gamma_{500}*\text{Phon} + \gamma_{600}*\text{Speed} + \gamma_{700}*\text{Lang} + \gamma_{1300}*\text{VWM}*\text{RCType} + \gamma_{1400}*\text{Inhib}*\text{RCType} + \gamma_{1500}*\text{Phon}*\text{RCType} + \gamma_{1600}*\text{Speed}*\text{RCType} + \gamma_{1700}*\text{Lang}*\text{RCType} + \gamma_{2300}*\text{VWM}*\text{QuestionType} + \gamma_{2400}*\text{Inhib}*\text{QuestionType} + \gamma_{1500}*\text{Phon}*\text{QuestionType} + \gamma_{2600}*\text{Speed}*\text{QuestionType} + \gamma_{2700}*\text{Lang}*\text{QuestionType} + \gamma_{12300}*\text{VWM}*\text{RCType}*\text{QuestionType} + \gamma_{12400}*\text{Inhib}*\text{RCType}*\text{QuestionType} + \gamma_{12500}*\text{Phon}*\text{RCType}*\text{QuestionType} + \gamma_{12600}*\text{Speed}*\text{RCType}*\text{QuestionType} + \gamma_{12700}*\text{Lang}*\text{RCType}*\text{QuestionType} + v_{00j} + v_{10j}*\text{RCType} + v_{20j}*\text{QuestionType} + v_{120j}*\text{RCType}*\text{QuestionType}$$

where Y_{ij} are the odds of subject i correctly responding to item j, γ_{000} is an intercept term representing grand mean accuracy, γ_{100}, γ_{200}, and γ_{1200} are the fixed effects of the experimental conditions, γ_{300} through γ_{700} are the fixed effects of the individual-difference composites on overall accuracy, γ_{1300} through γ_{1700} are the fixed effects of the individual-difference composites on the extraction-type effect, γ_{2300} through γ_{2700} are the fixed effects of the individual-difference composites on the question-type effect, γ_{12300} through γ_{12700} are the fixed effects of the individual-difference composites on the extraction type x question type interaction, u_{0i0} is the error for the subject intercept for subject i, u_{1i0}, $u_{2i0,}$ and u_{12i0} are the error terms for the random subject slopes for the experimental conditions for subject i, v_{00j} is the error for the item intercept for item j, and v_{10j}, v_{20j}, and v_{120j} are the error terms for the random slopes for the experimental conditions for item j, and ε_{ij} is the error for subject i on item j.

Responses to the attachment ambiguity items were modeled as:

$$\log(Y_{ij}) = \gamma_{000} + \gamma_{100}*\text{QuestionType} + u_{0i0} + u_{1i0}*\text{QuestionType} + \gamma_{200}*\text{VWM} + \gamma_{300}*\text{Inhib} + \gamma_{400}*\text{Phon} + \gamma_{500}*\text{Speed} + \gamma_{600}*\text{Lang} + \gamma_{1200}*\text{VWM}*\text{QuestionType} + \gamma_{1300}*\text{Inhib}*\text{QuestionType} + \gamma_{1400}*\text{Phon}*\text{QuestionType} + \gamma_{1500}*\text{Speed}*\text{QuestionType} + \gamma_{1600}*\text{Lang}*\text{QuestionType} + v_{00j} + v_{10j}*\text{QuestionType}$$

where Y_{ij} are the odds of subject i answering *yes* to item j, γ_{000} is an intercept term representing the grand mean of answering *yes* (response bias), γ_{100} is the fixed effect of the question type condition (low- or high-

attachment) on *yes* responses (sensitivity), γ_{200} through γ_{600} are the fixed effects of the individual-difference composites on response bias, γ_{1200} through γ_{1600} are the fixed effects of the individual-difference composites on sensitivity to the question type, u_{0i0} is the error for the subject intercept for subject *i*, u_{1i0} is the error term for the random subject slope of condition for subject *i*, v_{00j} is the error for the item intercept for item *j*, and v_{10j} is the error term for the random slope of condition for item *j*.