A new argument for co-active parses during language comprehension

Brian Dillon, Caroline Andrews, and Caren M. Rotello

University of Massachusetts, Amherst

Matthew Wagers

University of California, Santa Cruz

Abstract

One perennially important question for theories of sentence comprehension is whether the human sentence processing mechanism is *parallel* (i.e. it simultaneously represents multiple syntactic analyses of linguistic input) or *serial* (i.e. it constructs only a single analysis at a time). Despite its centrality, this question has proven difficult to address for both theoretical and methodological reasons (Gibson & Pearlmutter, 2000; Lewis, 2000). In the present study, we reassess this question from a novel perspective. We investigated the well-known *ambiguity advantage effect* (Traxler, Pickering & Clifton, 1998) in a speeded acceptability judgment task. We adopted a Signal Detection Theoretic approach to these data, with the goal of determining whether speeded judgment responses were conditioned on one or multiple syntactic analyses. To link these results to incremental parsing models, we developed formal models to quantitatively evaluate how serial and parallel parsing models should impact perceived sentence acceptability in our task. Our results suggest that speeded acceptability judgments are jointly conditioned on multiple parses of the input, a finding that is overall more consistent with parallel parsing models than serial models. Our study thus provides a new, psychophysical argument for co-active parses during language comprehension.

A new argument for co-active parses during language comprehension

Does the human parser maintain multiple co-active representations of a syntactically ambiguous sentence? This question has preoccupied many theories of human language processing because it speaks directly to the question of how syntactic information is represented during incremental sentence comprehension (Clifton & Frazier, 1996; Lewis, 2000).

One class of models—*serial* parsing models[1]—holds that humans can maintain only a single active syntactic description of the linguistic input at any point in time (Frazier, 1979; Frazier & Fodor, 1978; Frazier & Clifton, 1996; Kimball, 1973; Lewis & Vasishth, 2005; McElree, 2006; van Dyke & Lewis, 2003). *Parallel* models, by contrast, allow comprehenders to maintain multiple, co-active syntactic descriptions of the input at once (Gibson, 1991; Hale, 2001; Levy, 2008; MacDonald, Pearlmutter & Seidenberg, 1994; Trueswell, Tanenhaus & Garnsey, 1994). This broad class of parsing models can

[1] The terms *serial* and *parallel* are widely used in this debate, but we adopt them reluctantly. The crucial question is how many different syntactic representations are actively being updated as the parser takes in new information, not whether the computations happen simultaneously (cf. Lewis, 2000, who prefers the terms *single-path* versus *multi-path*).

be further divided into at least two distinct subclasses[2]. One group of models holds that comprehenders maintain multiple, distinct symbolic representations of the input (Gibson, 1991; Gorrell, 1995; Hale, 2001; Levy, 2008). A second class of parallel models holds that multiple syntactic analyses may be co-represented, but are merged into a single representation that blends or superimposes the constituent representations (Cho, Goldrick & Smolensky, 2017; Christianson, Hollingworth, Halliwell, & Ferreira, 2001; Ferreira & Bailey, 2004; Rasmussen & Schuler, 2017; Tabor & Hutchins, 2004; van der Velder & De Kamps, 2006; Vosse & Kempen, 2000, 2009).

This broad distinction is not only about how many mental representations there are at once, because the two models also differ in terms of the basic operations and memory characteristics of the parser: in particular, parallel parsers require either different elementary operations, like copying at choice points, or more sophisticated representations that allow substructures to be shared, like a chart (Earley, 1970). Parses in some grammatical theories are more easily represented with such structure-sharing mechanisms than in others (Lewis, 1998; Winograd, 1983). The choice also implicates different procedures of reanalysis (Lewis, 1998): backtracking or repair for serial parsers (e.g., Abney, 1989, Fodor & Inoue, 1998); selection, re-ranking or pruning of alternatives for parallel ones (e.g., Gibson, 1991; Jurafsky, 1996); and refinement of commitments for underspecified representations (e.g., Frazier & Clifton, 1996; Weinberg, 1993). Thus, finding evidence that could choose between serial and parallel theories has clear and

[2] We are grateful to Whitney Tabor for suggesting this classification to us.

definite implications for our cognitive models of how grammatical information is represented and used in real-time.

In this paper, we ask whether comprehenders can concurrently represent multiple analyses of a single sentence by investigating at end-of-sentence acceptability judgments. In our approach, we used speeded acceptability judgments rendered at or immediately following a potentially disambiguating word. We reasoned that such judgments should reflect the difficulty of integrating this input with the syntactic representations constructed by the comprehender prior to encountering that critical input (*i.e.* the parser's state at that point in incremental processing). We focus on the *ambiguity advantage effect* (AAE), the ironic finding that syntactically ambiguous sentences are easier to process than unambiguous sentences.

To preview our results and conclusions: we show that the AAE holds in speeded acceptability judgment measures, and use a Signal Detection Theoretic analysis to show that the trial-to-trial variance is comparable for both ambiguous and unambiguous sentences. Furthermore, we derive explicit models to predict the performance in our task under serial and parallel parsing models. We find strong evidence that parallel parsing models better predict task performance than serial models. Our results suggest that acceptability judgments are conditioned on multiple syntactic analyses simultaneously. On the assumption that speeded judgments reflect the parsing processes necessary to integrate the input into the parser's active syntactic representation(s), this finding is difficult to reconcile with parsing models that only assign a single, determinate analysis to ambiguous phrases.

The Ambiguity Advantage Effect

The ambiguity advantage effect (AAE) is the finding that globally ambiguous phrases are easier to process than lexically matched but unambiguous phrases. The AAE has mostly been demonstrated for syntactic adjuncts: those phrases that characteristically modify other parts of sentences but which are largely optional. To our knowledge, it has only been investigated in reading time measures. The first key study, Traxler, Pickering & Clifton (1998) used eye-tracking-while-reading and found that ambiguous sentences were read more quickly than unambiguous sentences. In their study, sentences were structured like (1), each containing a relative clause (RC; underscored) and two possible noun attachment sites (bold). In (1a), both attachment sites can be plausibly modified by the relative clause (RC) and the sentence is globally ambiguous. By contrast, (1b) and (1c) are effectively disambiguated by the predicate 'had a mustache': in (1b), the relative clause must attach 'low' to the recent noun phrase (NP) *the driver;* and in (1c), it must attach 'high' to the more distant NP.

(1) Three RC attachment configurations (Traxler, Pickering & Clifton, 1998)
   a. The **son** of the **driver** <u>that had a moustache</u> was really cool. [**AMBIG**]
   b. The car of the **driver** <u>that had a moustache</u> was really cool. [**LOW**]
   c. The **driver** of the car <u>that had a moustache</u> was really cool. [**HIGH**]

The finding that readers were faster in (1a) compared to (1b-c) implies that there is a processing cost associated with both high and low unambiguous RC attachments in this experiment. Similar findings have been observed for RC attachments with different disambiguating cues (Swets, Desmet, Clifton & Ferreira, 2008; Traxler et al., 1998), verb phrase (VP)/NP attachment ambiguities (van Gompel, Pickering & Traxler, 2001), reduced RCs (van Gompel, Pickering, Pearson & Liversedge, 2005), VP attachment

ambiguities (van Gompel et al., 2005), and even pronominal ambiguities (Grant, Dillon & Sloggett, 2014; Stewart, Holler & Kidd, 2007).

The AAE is a robust but seemingly counterintuitive effect, and it is a benchmark finding for models of incremental syntactic processing to explain. Traxler, Pickering & Clifton (1998) presented it as evidence against constraint-satisfaction models, which proposed that co-active alternative parses competed via lateral inhibition (e.g. MacDonald, Pearlmutter & Seidenberg, 1994; Vosse & Kempen, 2000). These models predict that, when multiple syntactic analyses compete for selection, the processor requires more time to settle into a single analysis. Because ambiguous material should trigger a competition, it should be associated with greater processing difficulty and higher reading times. The AAE seemed to disconfirm this prediction (Clifton & Staub, 2008; but cf. Green & Mitchell, 2006; Vosse & Kempen, 2009).

Van Gompel and colleagues (2000) proposed to explain the AAE with the *Unrestricted Race Model* (URM). The URM is a stochastic, serial parsing model in which a single attachment site is probabilistically selected for an ambiguous phrase, on every parse. This predicts that globally ambiguous conditions are easy to process, because for any attachment the resulting parse will be syntactically well-formed and plausible. In contrast, unambiguous conditions are difficult because, on some proportion of trials, the wrong attachment site is initially chosen and a time-consuming reanalysis is triggered on those trials. Crucially, "variable choice" models of this sort predict that this difficulty in unambiguous sentences only arises on a subset of trials where reanalysis occurs. On trials where the correct attachment is made, there should be no appreciable difficulty. Put differently, unambiguous sentences should thus either be very difficult to

process (when the 'wrong' attachment is chosen, and reanalysis is triggered) or no different from the ambiguous baseline (when the 'right' attachment is selected). This parsing model predicts a mixture distribution of processing difficulty within an experimental setting: there should be separate subgroups of 'easy' trials and 'hard' trials.

However, not all explanations of the AAE rely on reanalysis in a stochastic, serial parser (Levy, 2008; Logačev & Vasishth, 2016a; Swets et al., 2008). Levy (2008) hypothesizes that, when there are two compatible attachment sites, there is a richer context for predicting upcoming words. The critical word, 'moustache' in (1), has a higher conditional probability in a globally ambiguous sentence and thus a lower surprisal value, which is linked to ease of processing. In contrast, this same word is relatively less predictable in either unambiguous condition because it is only compatible with one RC attachment (Levy, 2008; p. 1156). Alternatively, Swets and colleagues (2008) offer an explanation in terms of underspecification (see also Ferreira & Patson, 2007; Frazier & Clifton, 1996). On this view, readers make time-consuming determinate attachments only when there is evidence for them (Swets et al., 2008). Still other interpretations are possible (see Logačev & Vasishth, 2016a). These models do not endorse a parser that obligatorily makes serial commitments, and so do not necessarily predict the same distinction between 'easy' and 'hard' trials that the URM does. Instead, they generally predict that there should be a cost paid for unambiguous sentences on every trial.

The AAE has been presented as one particularly clear piece of empirical support for models that invoke serial syntactic analysis (Clifton & Staub, 2008). This makes it a good starting point for addressing the question of whether the parser can concurrently

maintain multiple, co-active syntactic analyses. As described above, different accounts of the AAE make distinct predictions about the how the penalty associated with unambiguous sentences should be distributed across trials in an experiment (see discussion in Levy, 2008). These differing empirical predictions provide the motivation for the present investigation, in which we ask whether the penalty for unambiguous sentences obtains only in a small subset of trials in an experiment, as predicted by serial parsing models, or if instead it affects all trials in an experiment to some degree.

An Ambiguity Advantage in End-of-sentence Acceptability Judgments?

Previous research on the AAE has focused on processing times. Here we offer a new methodological approach in which we directly sample the shape of the underlying cognitive variable by analyzing speeded acceptability judgments in a signal detection theoretic framework. Addressing our distributional question using reading time measures poses serious methodological challenges. For example, Gibson and Pearlmutter (2000) observe that distinguishing two discrete modes in reading time data is difficult because of the shape of the distributions involved will tend to obscure this bimodality. Moreover, reading experiments typically do not yield enough data to clearly identify bimodality (however, see Vasishth & Nicenboim, 2016; Nicenboim & Vasishth, 2018 for a recent approach to this issue). In response to this difficulty, other researchers have pursued alternative response measures to test the question of bimodality in response behavior; for example, Farmer and colleagues (2007) used mouse-tracking to address the question of bimodality in response behavior. In the present work, we address our question using

acceptability judgment data. Acceptability judgments have been shown to readily reveal bimodality in response behavior (Dillon, Staub, Levy & Clifton, 2017).

Furthermore, using speeded acceptability judgments allows us to determine whether the AAE extends to dependent measures that do not directly reflect the incremental processing of the critical phrase. There is good reason to expect that it should. A number of studies have shown that sentence-medial parsing difficulty negatively affects end-of-sentence acceptability judgments in both speeded and unspeeded tasks (Ferreira & Henderson, 1991; Frazier & Clifton, 1998; Henderson & Ferreira, 1993; Tabor & Hutchins, 2004; van Dyke & Lewis, 2003; Warner & Glass, 1987). This observation receives a natural explanation under a serial parsing model: Ferreira and Henderson (1991) point out that reanalysis in a serial parser may simply not succeed on any given trial. If reanalysis fails, then the failure to find a grammatical parse should make the sentence appear ungrammatical. However, it is also possible that incremental parsing failures (Ferreira & Henderson, 1991; Vasishth, Bruessow, Drenhaus & Lewis, 2008) or the temporary creation of an ungrammatical representation (Sprouse, 2008; see also Clifton & Frazier, 2010; Fanselow & Frisch, 2006) create a persistent perception of sentence unacceptability that is reflected in end-of-sentence judgment measures, even if a successful parse is eventually found. In either scenario, the distribution of acceptability judgments reflects incremental parsing difficulty. It therefore stands to reason that the AAE will be evident in offline judgment measures: globally ambiguous sentences should be perceived as more acceptable than unambiguous sentences, because the latter will involve sporadic reanalysis or difficulty that in turn reduces their perceived acceptability.

Two Acceptability Judgment Models to Distinguish Serial and Parallel Parsers

If the AAE does extend to acceptability judgment measures, then the distributional predictions made by serial and parallel parsers map in a fairly straightforward way onto those measures. Just as other researchers have asked whether multiple parses co-determine reading times (Gibson & Pearlmutter, 2000), we ask: does the grammaticality of *two* parses contribute to the percept of acceptability on a single judgment trial, as would be expected under a parallel parsing model? Or does the grammaticality of only one parse determine acceptability on a trial-by-trial basis, as expected in a serial model? To address these questions, we first describe a formal model of the judgment process under both serial and parallel models.

We begin by assuming that any sentence can be mapped to a one-dimensional real number called *Acceptability*. This number is variable, and its precise value is affected by grammatical factors and by performance factors, such as: can a grammatical parse be found? how frequent are the words and constructions it contains? how efficiently is memory managed during parsing? etc. To address how acceptability is generated, the examples in (2) contrast sentences with grammatical and ungrammatical subject-verb agreement in an RC. Assume that sentences with grammatical subject-verb agreement correspond to a normal distribution over *Acceptability* whose mean is higher than sentences with ungrammatical subject-verb agreement. In sentences with RC attachment ambiguities, there are two possible subject-verb agreement relationships. How the RC is attached thus determines the Acceptability value. For (2a), we see that regardless of how the RC is attached, the resulting Acceptability value must be high because both

grammatically-possible subjects ('cousin', 'painter') have the appropriate features to agree with the verb ('knits'). And regardless of how the RC in (2b) is attached, the resulting Acceptability must be low because both otherwise grammatically-possible subjects ('cousins', 'painters') fail to agree with the verb. We refer to the sentences in (2) as "Pure" grammatical or "Pure" ungrammatical sentences. Figure 1, Panel A represents this situation graphically: the overall distribution of acceptability to the pure grammatical sentences (MULTIMATCH) is greater than that of the pure ungrammatical sentences (NOMATCH).

(2) **Pure grammatical or pure ungrammatical sentences (ambiguous attachment)**
a. Armand spotted the cousin of the painter who knits. **MULTIMATCH**
b. Armand spotted the cousins of the painters who knits. **NOMATCH**

(3) **Mixed grammatical/ungrammatical sentences (unambiguous attachment)**
a. Armand spotted the cousins of the painter who knits. **LOWMATCH**
b. Armand spotted the cousin of the painters who knits. **HIGHMATCH**

The unambiguous sentences (3), by contrast, have one grammatically-possible subject that agrees with the verb and one otherwise grammatically-possible subject that does not. We refer to the sentences in (3) as "Mixed" grammatical/ungrammatical sentences. How should the *Acceptability* for these Mixed sentences be distributed? The answer depends on whether the parser can represent only a single syntactic analysis of the input – i.e. is serial – or whether it can represent multiple analyses simultaneously – i.e. is parallel.

[FIGURE 1 GOES HERE]

Serial Parsing: A Formal Model of Acceptability

For a serial parser, the distribution of *Acceptability* for unambiguous Mixed sentences is a mixture distribution. On a given trial, either the high or low parse is initially computed, and in some of these trials the chosen parse will turn out to be contradicted by the agreement features on the RC verb. Accordingly, the resulting *Acceptability* value will be drawn from either the grammatical or ungrammatical distribution. Put differently, on any given Mixed trial, a serial parser will discretely sample from either the grammatical or the ungrammatical distribution at the end-of-sentence judgment depending on whether the correct attachment site was chosen. The resulting *Acceptability* distribution, then, is a mixture distribution. The properties of this mixture distribution, such as its variance, or whether it has multiple modes, are a function of i) the component distributions and ii) the probability of sampling from one distribution or another. More formally, we can say that the *Acceptability* distribution associated with grammatical and ungrammatical sentences is a normally distributed unidimensional random variable; for ease of presentation we refer to this as x. Without loss of generality, we assume that for ungrammatical sentences, x is normally distributed with a mean of 0 and a variance of 1 (the standard normal); for grammatical sentences, it is drawn from a normal distribution with its own mean and variance[3]:

[3] This is because the specific values of the mean and variance do not themselves matter, but rather the ratio of the variances for the grammatical and ungrammatical distributions. This ratio is measured empirically, and one distribution is set to mathematically convenient parameters to ensure model identifiability.

(4) $p_{UNGRAM}(x)\colon x \sim N(0,1)$

(5) $p_{GRAM}(x)\colon x \sim N(\mu_{gram}, \sigma^2_{gram})$

Let π_{high} represent the probability of choosing a high attachment on any trial. Given π_{high}, and the two component distributions in (4) and (5), we derive predicted *Acceptability* distributions for both Pure (2) and Mixed (3) sentences as follows:

(6) $p_{MULTIMATCH}(x) = \pi_{high}p_{GRAM}(x) + (1-\pi_{high})p_{GRAM}(x) = p_{GRAM}(x)$

(7) $p_{NOMATCH}(x) = \pi_{high}p_{UNGRAM}(x) + (1-\pi_{high})p_{UNGRAM}(x) = p_{UNGRAM}(x)$

(8) $p_{LOWMATCH}(x) = \pi_{high}p_{UNGRAM}(x) + (1-\pi_{high})p_{GRAM}(x)$

(9) $p_{HIGHMATCH}(x) = \pi_{high}p_{GRAM}(x) + (1-\pi_{high})p_{UNGRAM}(x)$

Figure 1, Panel B represents graphically the distribution of *Acceptability* under a serial parsing model, derived from the equations in (6)-(9). Note that although the predictions of serial models are sometimes cast in terms of 'bimodality' (e.g. Levy, 2008; Staub & Clifton, 2008), from this perspective bimodality is not strictly speaking diagnostic of a serial parser. Instead, what matters is that the overall distribution is a mixture that consists of two distinct component distributions. In the aggregate, this distribution may appear unimodal. However, it will crucially have greater variance than either of its component distributions.

Parallel parsing: a formal model of acceptability

In contrast, for a parallel parser the *Acceptability* distribution for Mixed sentences is derived differently. We model a generic parallel parsing model that maintains multiple, weighted syntactic descriptions of the input at once (e.g. Gibson, 1991; Jurafsky, 1996; MacDonald et al., 1994; Tabor & Hutchins, 2004; Vosse & Kempen, 2000). To derive the *Acceptability* distribution under such a model, we assume that an acceptability judgment is conditioned on all active parses which are active at the decision point, weighted by their strength or probability. In the present case, this means that both the acceptability of the high attachment and the low attachment codetermine the final *Acceptability* value, for Pure and Mixed sentences alike. More formally, the distribution of *Acceptability* for Mixed sentences on the parallel model is a random variable that is itself a weighted sum of two random variables: the *Acceptability* value of the high parse and that of the low parse. Let *x* again represent *Acceptability*, the random variable of interest. As above, we assume that the distribution of this variable associated with simple grammatical agreement and ungrammatical agreement parses is a normally distributed unidimensional random variable *x*, constrained to the standard normal for ungrammatical sentences:

(10) $p_{UNGRAM}(x)\colon x \sim N(0,1)$

(11) $p_{GRAM}(x)\colon x \sim N(\mu_{gram}, \sigma^2_{gram})$

Let π_{high} represent the weight associated with the high attachment parse; the weight associated with the low attachment parse is constrained to (1-π_{high}). These weights are constrained to be positive. Under these conditions, we may again derive predicted *Acceptability* distributions for the Pure (2) and Mixed (3) sentences alike. For all four conditions, the *Acceptability* value is a normally distributed random variable that reflects the weighted sum of the component distributions of the high and low parses[4]:

(12) $p_{MULTIMATCH}(x)$: $\mathcal{N}(\mu_{gram}, \pi_{high}^2\sigma_{gram}^2 + (1-\pi_{high})^2\sigma_{gram}^2)$

(13) $p_{NOMATCH}(x)$: $\mathcal{N}(0, \pi_{high}^2 + \left(1-\pi_{high}\right)^2)$

(14) $p_{LOWMATCHH}(x)$: $\mathcal{N}((1-\pi_{high})\mu_{gram}, \pi_{high}^2 + (1-\pi_{high})^2\sigma_{gram}^2)$

(15) $p_{HIGHMATCH}(x)$: $\mathcal{N}(\pi_{high}\mu_{gram}, \pi_{high}^2\sigma_{gram}^2 + \left(1-\pi_{high}\right)^2)$

Figure 1, Panel C represents graphically the distribution of *Acceptability* under a parallel parsing model, derived from the equations in (12)-(15). If the underlying distribution of *Acceptability* is normal, the resulting distribution is also a normal

---

[4] The mean of the weighted sum of *n* normally distributed random variables is itself a random variable with mean $\sum_{i=1}^{n}\lambda_i\mu_i$, where λ_i is the weight for each variable in the sum. The variance of this random variable is $\sum_{i=1}^{n}\lambda_i^2\sigma_i^2$. The values of the mean and variance in (12)-(15) are derived from these.

distribution, with a mean and variance that that can be analytically derived from the mean and variance of the component distributions as in (12)-(15).[5]

With these models in hand, we are in a position to distinguish serial and parallel models of judgment performance on the Mixed sentences: if we find evidence that the distribution of *Acceptability* in Mixed sentences is more similar to a mixture model, then serial parsing models are supported. In contrast, if the distribution of *Acceptability* in Mixed sentences appears to be an admixture or blend, then parallel models are supported. At present, our predictions are stated in terms of an unobservable cognitive variable, *Acceptability*. In order to evaluate these predictions, we turn to Signal Detection Theory (SDT) to have a way of measuring this underlying cognitive variable.

Signal Detection Theoretic Analysis of Acceptability

To judge a sentence's acceptability is to make a decision under uncertainty. From the perspective described above, this process is understood as a mapping from a noisy cognitive signal about the well-formedness of a sentence (i.e. *Acceptability*) to one of several response options offered to a participant (e.g. binary *yes-no* acceptability judgments, *n*-point Likert scales, etc.). We can analyze this process in SDT (Green & Swets, 1966; Macmillan & Creelman, 2005; cf. Bader & Haüssler, 2010, for a related attempt to analyze acceptability judgments with signal detection theory).

---

[5] Code for all analyses, models, and experimental data presented in this paper can be found at: https://osf.io/sd3hu/.

SDT makes explicit the role of decision processes, and describes the balance of possible decision errors (e.g., *missed* targets and *false alarms* to lures). A key advantage of SDT is that it allows independent estimates of the accuracy with which classes of stimuli can be discriminated and the tendency for raters to prefer to respond with one response class or another (i.e., their response biases).

[FIGURE 2 GOES HERE]

Figure 2 (left panel) shows an SDT model for a simple judgment task involving two classes of stimuli, such as the pure grammatical and pure ungrammatical sentences tested above. The *x*-axis here represents the *evidence* that an observer uses to make a decision. In the context of an acceptability judgment task, the evidence corresponds to *Acceptability* defined above (Bader & Häussler, 2010). A participant in a binary acceptability judgment task is assumed to set some (potentially arbitrary) decision *criterion* (black vertical line), above which a positive response (e.g. "grammatical") is offered. The assumption of an arbitrary criterion setting is how SDT represents the idiosyncratic interindividual differences in how participants make use of the response options given in a judgment task (*scale bias*; Schütze & Sprouse, 2014). The area under the Target distribution above the criterion is the proportion of Targets that elicit a correct response, namely the *hit rate (H)*. The area under the Lure distribution above the decision criterion provides the *false alarm rate (F)*.

In sum, an SDT model of acceptability judgments endorses the view that acceptability judgments are based on a noisy, continuous signal of a sentence's

wellformedness. By assuming arbitrary response criterion placement over this signal, SDT offers a model for how participants map this signal onto discrete response options in an experiment, and offers an explanation of why participants vary in their use of the response scale offered in an acceptability judgment task (Schütze & Sprouse, 2014).

From this theoretical perspective, we can derive a measure of the discriminability or psychological distance between two classes of stimuli in d_a, a measure of the distance between the means of the distributions in units of their root mean squared standard deviation. When there are multiple response criteria in an experimental setting, the resulting (*F*, *H*) pairs yield a theoretical curve called a *receiver operating characteristic* (ROC; Fig. 2, middle panel); transforming both *F* and *H* to their *z*-score equivalents yields a *zROC* (Fig. 2, right panel). A theoretical ROC curve connects all pairs of hit and false alarm rates that reflect the same level of discrimination accuracy as measured by a specific summary statistic, such as *d'* or percent correct. Each point on the ROC reflects the same accuracy, according to that measure, but a different response bias. These different biases can result from independent experimental conditions in which a more conservative or liberal response bias is induced and participants make a binary (yes/no) decision, or from responses associated with different confidence levels within a single condition: the same ROC will result (e.g., Egan, Shulman, & Greenberg, 1959; Dube & Rotello, 2012). Smaller values of *H* and *F* indicate a conservative response bias (i.e., a preference for a "no" response); these points occur toward the lower-left end of the ROC. Larger values of *H* and *F* indicate a liberal bias (i.e., preference for "yes"); the operating points occur at the upper-right end of the ROC. Higher values of *H* relative to *F* reflect higher decision accuracy.

Importantly for present purposes, the shape of the (*z*)ROC gives us clues as to the underlying distribution of the *Acceptability* values: the slope of the *z*ROC equals the ratio of the standard deviations of the Lure to the Target distribution (i.e., 1/*s*, where *s* = Target distribution standard deviation). For example, when the Target distribution is more variable than the Lure distribution, the slope of the zROC is less than 1, a result that is ubiquitous in the recognition memory literature (e.g., Ratcliff, Sheu, & Gronlund, 1992). Moreover, it is possible to directly observe mixture distributions in the shape of the *z*ROC: deCarlo (2002) shows that mixture distributions of evidence appear as curvilinearity in the *z*ROC that increases with the difference between the means of the latent distributions.

Experiment

We conducted a speeded acceptability judgment experiment to test our primary question of interest: is the distribution of *Acceptability* associated with unambiguous attachment sentences better characterized as a mixture, as predicted by serial models, or as a blend, as predicted by parallel models? To evaluate this, we tested sentences in the four conditions presented in (3) and (4) above. We had three goals.

First, we sought to replicate the ambiguity advantage effect in standard dependent measures for speeded acceptability tasks, such as judgment accuracy and judgment RTs. In particular, we expected that pure grammatical and ungrammatical sentences would be classified more accurately, and more quickly, than their mixed grammaticality counterparts. This would extend the AAE to a novel dependent measure. Second, we analyzed those judgments in an SDT framework to evaluate whether acceptability

judgments yield data patterns consistent with the modeling framework described above. Third, and most critically, we directly fit the serial and parallel parsing models above to the data we collected. In doing so, we sought to predict performance on the mixed conditions on the basis of performance in the pure conditions.

Method

Participants

Eighty-one participants were recruited to participate in the experiment. Of these, 34 were undergraduates at UMass Amherst and 47 were undergraduates at UC Santa Cruz. Participants gave informed consent and were compensated with course credit for their participation. All participants were native speakers of American English, over 18 years old. The experimental protocol was approved by the Institutional Review Boards at both UMass Amherst and UC Santa Cruz.

Materials

Forty critical item sets were developed. These critical materials were distributed into four Latin Squared lists, and each list was combined with the same set of sixty-eight fillers. Each participant was randomly assigned to a list. Forty-four of the filler items were grammatical. Thus, across the experiment the ratio of grammatical to ungrammatical sentences was 1:1. Filler sentences were designed to prevent participants from deploying any superficial 'scanning' strategies to make judgments, and included errors with unambiguous relative clause attachments, agreement with coordinated

subjects, reflexive agreement, number sensitive items (e.g. *together*), and implausible subject-verb combinations (e.g. *the car fainted*).

Across items, there was some variability in the position of the critical agreeing verb relative to the end of the sentence. The verb occurred sentence finally (e.g. ... *who* ***knits***.) in nineteen out of forty items (47.5%). There was a single word between the inflected verb and the end of the sentence (e.g. … *who* ***was*** *blogging.*) in sixteen out of nineteen items (40%), and in the remaining five items (12.5%) the end of the sentence occurred two words after the inflected verb (e.g. … *who* ***was*** *on TV.*). The full list of experimental stimuli is available at https://osf.io/sd3hu.

Prior to the experiment, we conducted an offline norming study to determine the attachment preferences in our experimental stimuli. 40 native speakers of American English were presented with the ambiguous, grammatical versions of our stimuli, and asked which NP in the sentence performed the action described in the relative clause. Overall, our items had a low attachment bias: participants interpreted the relative clause high on only 30% of trials. This value ranged from a minimum of 10% to a maximum of 55% across individual items, and half of the items had a high attachment bias between 22% and 38%. The attachment biases measured in the norming study may be seen as an empirical estimate for the parameter π_{high}.

Procedure

Stimuli were displayed and responses were collected using the Linger software (Rohde, 2003). Each experimental trial began with a fixation cross in the center of the screen for 1s. Afterwards, a sentence was presented one word at a time. Each word appeared for 225ms followed by 100ms of blank screen. At the end of each trial, participants were prompted to make a binary acceptability judgment decision, with *f* indicating 'grammatical' and *j* indicating 'ungrammatical.' Participants were given a 2-second deadline for the binary acceptability judgment. Following each judgment, participants rated their confidence in their decision on a three-point scale using the number keys 1-3. On each trial, the ends of the confidence scale were labeled, ranging from 'not at all confident' (1) to 'very confident' (3). Participants were under no time pressure to offer their confidence rating.

An experimental session began with an informed consent form and a demographic survey. Following that, the participant was told that they would be reading sentences one at a time and that they would be judging whether each one sounded like 'natural, grammatical English.' The experimenter briefly explained what was meant by grammatical (colloquial, idiomatic English). Several practice trials were given, including practice trials with correct and incorrect agreement.

Regression analysis

We used mixed effects regression models to analyze i) percentage *yes* responses on the binary judgment, ii) reaction times to the binary judgment, and iii) confidence ratings. For all models, we used the maximal random effects structure (Barr, Levy, Scheepers & Levy, 2013) where possible. For binary judgment responses, logistic

regression was used to model the probability of offering a 'grammatical' response. For confidence ratings, ordinal regression was used on the restricted three-point confidence, ranging from 1 (not at all confident) to 3 (very confident). All reaction times were log-transformed prior to analysis to ensure normally distributed residuals for linear models. The following fixed-effects contrasts were used: *ambiguity (*0.25 for LOWMATCH, HIGHMATCH and -0.25 for MULTIMATCH, NOMATCH), *height* (0.5 for HIGHMATCH, -0.5 for LOWMATCH, 0 otherwise) and *grammaticality* (0.5 for MULTIMATCH, -0.5 for NOMATCH, 0 otherwise). For reaction time and confidence ratings models, we further used *accuracy* as a fixed effect predictor (0.5 for *incorrect response*, -0.5 for *correct response*; 'grammatical' responses were accurate for all conditions except NOMATCH), and tested for interactions of response with the other fixed effect predictors. For linear models, we accepted coefficients with *t*-values greater than 2 as significant (Gelman & Hill, 2007).

Prior to all analyses, we rejected any trial on which a participant did not make a response prior to the deadline for the binary judgment. This resulted in the rejection of 299 trials (3% of data overall).

ROC analysis

For empirical ROC analysis, we combined the binary judgment and the confidence rating into a six point response scale ranging from *very confident grammatical* (1) to *very confident ungrammatical* (6). Prior to ROC analysis, we aggregated responses across all participants (Macmillan & Kaplan, 1985). On the aggregated data, we calculated the hit rate and false alarm rate at each confidence level (Macmillan &

Creelman, 2005). We performed three comparisons, allowing each of MULTIMATCH (the Pure grammatical condition), LOWMATCH, and HIGHMATCH (the two Mixed conditions) to contribute the hit rate for one comparison. For each, the NOMATCH condition determined the false alarm rate. In the interest of brevity, we refer to these comparisons simply as the MULTIMATCH, LOWMATCH, and HIGHMATCH comparisons.

For each of these three comparisons, we fit unequal variance signal detection (UVSDT) Models to estimate both accuracy (measured in d_a) and slope ($1/s$) for all three comparisons. We estimated these parameters for each of the three comparisons using a bootstrap procedure, drawing 500 bootstrap samples by sampling (with replacement) at the level of experimental participant. The goal of this analysis was twofold. First, we aimed to determine whether the SDT assumptions about the underlying evidence distributions were well met for acceptability judgment data. Second, we wished to provide a theory neutral measurement of the relative amount of variation in the judgment distributions across the three comparisons.

We also directly fit the serial and parallel parsing models of Acceptability, described in the introduction, to our data. We attempted to predict performance on the unambiguous conditions on the basis of performance on the ambiguous conditions. For both serial and parallel models, this was done by positing a single free parameter π_{high}. The *serial* model modeled the performance on the unambiguous conditions as the result of a mixture process: on any given trial, the perceiver discretely creates a high or low attachment parse, which in turn results in a sample from the acceptable or unacceptable distribution, depending on condition. In this model, π_{high} is interpreted as the probability

of creating a high attachment on any given trial. The *parallel* model holds that the perception of acceptability on any given trial is a weighted sum of the acceptability of the high and low parses. In this model, π_{high} is interpreted as the relative weight given to the high parse. The *Acceptability* distributions for each model were combined with *(k-1)* free criterion locations ($c_1, \ldots, c_{k-1}$) to derive predicted response proportions in each of k response categories. For models reported in this paper, $k = 6$, yielding a total of 8 free parameters for both models (the mixture parameter, 5 criterion locations, plus the mean and standard deviation of the pure grammatical distribution). Models were fit using R (R Core team) by minimizing $-2\mathcal{L}(\theta)$ for the model fit to response proportions in each of the 6 response categories in our 4 experimental conditions. In order to estimate how reliable observed $-2\mathcal{L}(\theta)$ differences were, we performed bootstrap analysis using the same parameters as above.

Results: RT and judgment analyses

Binary acceptability responses

Descriptive statistics for each of the judgment measures, and results from the regression analyses over those measures, are presented in Tables 1 and 2. Logistic mixed-effects regression analysis revealed significant effects of all fixed-effects predictors. Participants were less likely to offer a 'grammatical' response for *NOMATCH* conditions compared to *MULTIMATCH* conditions, and were less likely to respond 'grammatical' to *HIGHMATCH* conditions than *LOWMATCH* conditions. Importantly, there was also an

overall effect of ambiguity: participants classified unambiguous sentences grammatical less often than ambiguous sentences.

[TABLE 1 GOES HERE]

[TABLE 2 GOES HERE]

Confidence ratings

Ordinal mixed-effects regression analysis reveals that raters were significantly less confident in incorrect responses (see Tables 1 and 2). We also observed a significant interaction of *Accuracy* and *Ambiguity*; this interaction reflected lower confidence ratings for correct unambiguous trials than correct ambiguous trials. Similarly, there were higher confidence ratings for incorrect unambiguous trials than incorrect ambiguous trials. Last, we observed an interaction of *Accuracy* and *Height*, driven by higher confidence ratings for correct responses to *LOWMATCH* trials.

Reaction times

Linear mixed-effects regression analysis of the log-transformed judgment times revealed several significant effects. First, reaction times for inaccurate trials were slower than for accurate trials. In addition, there was a significant main effect of ambiguity, such that unambiguous trials were judged significantly more slowly than ambiguous trials. However, this effect was qualified by an interaction of *Ambiguity* and *Accuracy* that mirrored pattern observed in confidence ratings: correct responses were significantly slower for unambiguous trials than for ambiguous trials. Last, we observed an interaction

of *Height* and *Accuracy,* such that correct responses had longer RTs for *HIGHMATCH* conditions than for *LOWMATCH* conditions.

Discussion: RT and judgment analyses

Judgments and confidence ratings reveal that there is an ambiguity advantage effect in the speeded acceptability judgments. Ambiguous *MULTIMATCH* conditions were correctly judged grammatical more often than either *LOWMATCH* or *HIGHMATCH*.

For accurate trials in ambiguous conditions—*MULTIMATCH* and *NOMATCH*—confidence was high and response times were fast, while confidence was low and RTs were long for inaccurate responses. The unambiguous *LOWMATCH* and *HIGHMATCH* conditions displayed a different pattern. For correct trials, confidence was lower and RTs were longer for unambiguous conditions compared to ambiguous conditions. For grammatical sentences, this replicates the ambiguity advantage effect in two further dependent measures: judgment latency and confidence ratings. For incorrect trials, confidence was higher and RTs were faster for unambiguous conditions compared to ambiguous conditions. This overall ambiguity advantage was again qualified by an effect of attachment height. In particular, we observed that an effect of the low bias both in confidence ratings and RTs: correct grammatical responses were issued with less confidence, and greater RTs, for *HIGHMATCH* conditions compared to *LOWMATCH* conditions.

In short, analysis of reaction times and confidence measures shows that ambiguous conditions were judged more quickly and more confidently than were unambiguous sentences.

Results: SDT analysis

UVSDT Analysis

Unequal variance SDT fits to our data are presented in Table 3. Receiver operating characteristics are plotted in Figure 3.

With respect to d_a measures of accuracy, we observe clear differences between all conditions. The *MULTIMATCH* comparison was associated with the highest accuracy, followed by *LOWMATCH*, and then *HIGHMATCH*. Insofar as the *MULTIMATCH* condition had higher accuracy than the unambiguous conditions, this pattern reflects an ambiguity advantage in judgment measures. However, our analysis also reveals a significant difference between *LOWMATCH* and *HIGHMATCH* conditions, with higher accuracy on unambiguous low attachment sentences.

Turning to 1/s, we observe that all three comparisons had values close to 1, suggesting equal variance between the grammatical conditions and the ungrammatical condition. While *LOWMATCH* and *MULTIMATCH* comparisons had nearly identical estimates of $1/s$[6], the estimate for *HIGHMATCH* was lower than either of these by .1; this

[6] Note that the lure distribution was identical for all comparisons, and for reasons of model identifiability, its variance was fixed to 1. This crucially allows direct

pattern suggests that the *HIGHMATCH* conditions had greater variance in judgments than either *LOWMATCH* or *MULTIMATCH* comparisons. When compared against *MULTIMATCH*, this difference was significant; when compared against *LOWMATCH* it was not. Thus while this analysis suggests that the variance in judgments is largely comparable across all three comparisons, we do find limited evidence for greater variability in the *HIGHMATCH* conditions relative to all other comparison.

[TABLE 3 GOES HERE]

[FIGURE 3 GOES HERE]

Serial versus parallel models

The best-fit serial mixture and parallel ad-mixture models are shown in Figure 4. Overall, the models fit the observed data fairly well; for both models, the overall predicted accuracy was slighter lower than observed. For the observed data, there was a clear advantage for parallel models over the serial models ($-2\mathcal{L}(\theta)_{\text{SERIAL}} - -2\mathcal{L}(\theta)_{\text{PARALLEL}} =$ 23.8; 95% CI based on bootstrap = [3.4, 45.0]). The parallel model consistently had lower $-2\mathcal{L}(\theta)$ values across bootstrap samples: 99% of resampled data sets were fit better by the parallel model.

---

comparison of the magnitude of 1/*s* across comparisons.

Inspection of Figure 4 suggests that the parallel model achieves lower BIC scores primarily because of its ability to fit the slope of the zROC line. The serial model does a poorer job of this, and consistently predicts shallower slopes than what are actually observed. This pattern suggests that the degree of variability observed in the unambiguous conditions is more consistent with a parallel model than a serial model. In other words, it appears that the variability in *Acceptability* in the Mixed conditions does not appear to be great enough to support the predictions of a serial parser.

One further pattern that is evident in Figure 4 is that both the parallel and serial models provide somewhat imperfect fits to the data. This suggests that some aspect of the data does not fit the generative model implied by each of these models. We see several possible reasons for this. First, we note that the zROCs are not perfectly linear: there is a bowing of the zROC function in the middle three points of the curve, a pattern that is most pronounced in the MULTIMATCH condition. Ratcliff, McKoon and Tindall (1994) note that this pattern may arise if there is noise in an experimental setting that results in a subset of decisions being uniformly distributed of the response. For example, this pattern may plausibly arise when there is a small subset of trials where a participant's attention lapses, leading them to offer responses randomly. Such a noise process is not explicitly represented in our models, and so could create a disconnect between the model fits and the empirical data. Second, we note that the best fit overall is the one that allows the model to fit response behavior in all three comparisons under highly restrictive conditions. In particular, our model assumes that the weights on the two parses in both models must sum to one. This is a natural constraint for the serial model, where the weight is interpreted as the probability of attaching at one of two possible attachment sites. For the parallel model, however, this constraint may not be warranted: on this model, it is theoretically possible to assign arbitrary weights to both the high and low attachment parses. It is possible that with this greater flexibility, the parallel model could achieve a more accurate fit overall.[7]

---

[7] Exploratory fits with such a model confirm that this is indeed the case. An

[FIGURE 4 GOES HERE]

Model mimicry analysis

The modeling results suggest a clear advantage for parallel models over serial models. In order to interpret this as evidence in favor of parallel models, it is important to determine whether the two models can, in principle, be distinguished using our experimental paradigm. To evaluate this, we performed a model mimicry analysis (Wagenmakers, Ratcliff, Gomez, & Iverson, 2004). We generated 550 bootstrap samples from our original data set. For each sample, we fit both a serial mixture and parallel admixture model. From each fitted model, we simulated a data set, creating one simulated data set generated from a parallel model and one from a serial model. We then fit both the serial and parallel models to each of these simulated data sets. If the serial models provide the better fit to the simulated serial data, and the parallel models the simulated

---

'unrestricted' parallel model that replaces the low attachment weight (1-π_{high}) with a totally free parameter π_{low} achieves a significantly better fit to the data ($-2\mathcal{L}(\theta)_{\text{PARALLEL:RESTRICTED}} - -2\mathcal{L}(\theta)_{\text{PARALLEL:UNRESTRICTED}} = 32.7$). However, this model has significantly more flexibility to fit the data than the restricted parallel model described in the text. Thus, in order to provide the fairest comparison to the serial model, where the constraints on parameter values are theoretically grounded, we choose to retain our focus in the text on the restricted version of the parallel model.

parallel data, then we conclude that the two models can be distinguished in this experimental design.

The results of this analysis are presented in Figure 5. As before, we evaluated fit using $-2\mathcal{L}(\theta)$. The optimal criterion for distinguishing the two models was a difference of 4.99 in $-2\mathcal{L}(\theta)$. Adopting this criterion, we found that more than 99% of generated data sets were better fit by the correct generating model. Although classification errors were very few, they were not equally distributed among serial and parallel models. Simulated data from the parallel model was more likely to be better fit by the serial model than the other way around. This is suggestive evidence that that of the two models, the serial model may have more flexibility in fitting possible data sets (it is somewhat more likely to falsely "claim" data as being self-generated). This bolsters confidence in the results reported above: the parallel model appears to be the more constrained model in terms of its ability to fit diverse data patterns, yet it consistently outperformed the serial model on the data we collected.

[FIGURE 5 GOES HERE]

General Discussion

We investigated the ambiguity advantage effect in a speeded binary acceptability judgment task, with a secondary confidence rating measure. We analyzed the results using both a traditional analysis of ratings and reaction times, as well as a secondary analysis using signal detection theory. In addition, we derived predictions about

Acceptability distributions using explicit serial and parallel parsing models. Our goal was to evaluate if both parses of an in-principle ambiguous relative clause co-determine sentence acceptability, or if instead, only one parse does.

In an analysis of responses and reaction times, we replicated the AAE in a speeded acceptability judgment measure: there were fewer correct grammatical responses to unambiguous sentences than ambiguous sentences, and participants were slower to accept unambiguous sentences than ambiguous sentences. In a secondary confidence rating measure, we observed that comprehenders were overall less confident in their judgments to unambiguous sentences. These empirical conclusions were further refined with a signal detection theoretic analysis. This analysis showed that the ambiguous MULTIMATCH condition had higher *Acceptability,* expressed in d_a, than either of the unambiguous sentences. In turn, we also observed higher d_a for LOWMATCH sentences than HIGHMATCH sentences. In measures of variance derived from the signal detection analysis, we failed to find any evidence that unambiguous HIGHMATCH conditions had higher variance in *Acceptability* than did ambiguous MULTIMATCH conditions. However, we did find evidence that suggests that the unambiguous HIGHMATCH comparisons did have more variance than either the MULTIMATCH or LOWMATCH comparisons. Last, we fit explicit serial and parallel models to our data. These models predicted the distribution of the Mixed conditions on the basis of performance in the Pure conditions, under both a serial and a parallel parsing model. This modeling exercise revealed clear evidence in favor of the parallel model: performance in the Mixed conditions was better predicted by a parallel parser, and this was true for 99% of bootstrapped data sets in our analysis.

Overall, across all measures and analyses, we found evidence that better accords with the view that both relative clause attachment sites are co-active until the moment of disambiguation, and that speeded acceptability judgments are conditioned on both attachments. For example, with respect to reaction times and confidence ratings, the pattern of responses suggests that participants were overall less sure of their responses in unambiguous conditions than they were for ambiguous conditions. This is consistent with a parallel parsing model that allows both parses to simultaneously contribute to the judgment process: in the Mixed conditions, there are conflicting acceptability signals resulting from the high and low attachments. This conflict creates uncertainty in the judgment process, lowering confidence ratings and increasing judgment times for the binary judgment. In a serial model, it is not clear how the acceptability of the parse not chosen can impact judgment confidence or reaction times; in a sense, serial models make the strong claim that Mixed sentences are always perceived as Pure sentences, because only one attachment is ever active. In brief, our serial model offers no mechanism to allow the parse not chosen to interfere in the judgment process; yet this is what the effect of ambiguity on judgment times and confidence ratings seems to suggest.

The signal detection theoretic analysis points to a similar conclusion. First, we note that the ROC functions we measured in our experiment appear remarkably well-behaved. The *z*ROCs appear largely linear, which is consistent with the modeling assumptions made in the input: it appears that speeded acceptability judgments are well modeled by a Gaussian signal detection process. Furthermore, the d_a accuracy measures indicate that ambiguous Pure sentences have greater *Acceptability* than unambiguous Mixed sentences; this replicates the AAE in judgment measures, and offers some

validation of our choice of dependent measure. Interestingly, we also observed that *1/s* was approximately 1 for the MULTIMATCH comparison. Importantly, this is consistent with the specific modeling assumptions we made in the introduction: namely, that the distribution of the Pure grammatical and Pure ungrammatical sentences are normal and of equal variance. *1/s* was also approximately 1 for the LOWMATCH comparison, indicating that the variance in the Mixed LOWMATCH condition is roughly equal to that of the Pure conditions; we thus find no evidence of bimodality, or increased variance, in the LOWMATCH conditions, contravening the predictions of a serial parsing model. The situation was slightly different for the mixed HIGHMATCH comparison. The *1/s* estimate for HIGHMATCH was slightly lower than either the LOWMATCH or MULTIMATCH comparisons, indicating that there was greater variability in judgments in this condition. However, even so, this variability was not large enough to satisfy the predictions of the serial parsing model.

Last, and perhaps most critically, this conclusion is directly supported by the head-to-head comparison of serial and parallel parsing models. We were able to predict performance in the Mixed conditions by assuming that *Acceptability* in these trials was a weighted sum of two, co-active parses that each contribute their own *Acceptability* in proportion to their weight. In our view, the parallel model's ability to provide fairly good predictions to performance on the mixed conditions on the basis of the pure conditions by adopting only a single free parameter constitutes an argument in favor of the parallel model.

To return to our original question: can the human parser maintain multiple co-active representations of a syntactically ambiguous sentence? Our data and modeling

suggest that the answer to this question is *yes*. It seems that both parses co-determine the acceptability of an unambiguous Mixed sentence, and they appear to do so on the vast majority of experimental trials. We find no evidence to support the view that there is substantial trial-to-trial variation in the *Acceptability* of our mixed sentences, a key prediction of serial parsing models.

Relating acceptability and incremental processing

Our primary theoretical conclusion is that the parser can simultaneously represent multiple analyses of a single sentence, and that this is responsible for our finding that multiple parses jointly contribute to the acceptability of a sentence. Our theoretical conclusion rests on the assumption of a fairly direct mapping between parser state and acceptability: acceptability is a straightforward function of whatever representations are active at the point when the judgment is formed.

This is one possible function that could link incremental parsing behavior and acceptability judgments, but it is not the only possibility[8]. It may be, for example, that acceptability judgments reflect the average acceptability of all parse states that the parser had created over the course of a sentence (Sprouse, 2008). Under this linking hypothesis, a serial, probabilistic parser that can rapidly reanalyze an ungrammatical parse might be able to explain our findings. This is because any reanalysis that occurs during incremental

---

[8] We are grateful to Whitney Tabor for suggesting this framing of the issue.

processing will negatively impact acceptability, reflecting both on the unacceptable pre-reanalysis parse and the acceptable post-reanalysis parse.

We cannot rule this possibility out categorically, but we believe it unlikely for the present data for two reasons. First, in our experiment, acceptability judgments were rendered at, or immediately following, the point of disambiguation (*i.e.* the sentence-final inflected verb). This leaves little time for reanalysis before the judgment is required; the speeded response demands on the participants plausibly creates a pressure to judge the sentence based on whatever parse they have constructed at the point that the critical word is encountered. Second, if participants were parsing serially and rapidly reanalyzing ill-formed input at the point when a judgment was demanded, then we might expect long response times to the ungrammatical NOMATCH condition. This is because participants would have to reanalyze and compute both high and low attachments in this condition in order to recognize that these sentences are unambiguously ill-formed. Such an exhaustive search would be a time-consuming computation. Yet there is no trace of this in our data: accurate 'no' responses to the ungrammatical condition were in fact the fastest responses observed in our experiment. This suggests that the evidence that the string was ill-formed was relatively clear or unambiguous, allowing rapid, accurate acceptability judgments (see Hammerly, Staub & Dillon, 2018, for a model of judgment RTs; see also Ratcliff & McKoon, 2008). In this sense, the fast and accurate 'no' responses to the NOMATCH condition are more consistent with a parallel parsing model.

Implications for models of parsing

One important question is how this finding links to the incremental processing measures of the AAE discussed in the introduction. If we assume that the AAE in

judgment measures and online processing measures reflect a single underlying processing mechanism, then we would expect that the AAE observed in reading measures is similarly gradient in nature. If true, this constitutes an argument against the URM, and other serial, stochastic models of this effect.

Caution is warranted in drawing too strong a parallel between the AAE in our task and that seen in online reading measures. It has been shown that the experimental task context can alter parsing behavior. In particular, the depth of syntactic analysis pursued may be modulated by task demands (Logacev & Vasishth, 2016a; Swets et al., 2008). Since comprehenders did not need to form a single, coherent interpretation of our stimuli for the purposes of answering comprehension questions, our task context may not have created a strong pressure to form a single parse of the input. If our judgment task is one where comprehenders need not form a determinate attachment, then this may be what drives the pattern of responses we see. In other words, a shallow, parallel representation of the input may be 'Good Enough' to render an acceptability judgment (Ferreira & Patson, 2007; Swets et al., 2008). In light of this, we cannot at present confidently generalize beyond our task to other task contexts, such as reading experiments, where forming a coherent interpretation is often necessary (Swets et al., 2008).

Indeed, if we do assume that the ambiguity advantage effect seen in online reading measures and the one observed in our experiment reflect the same phenomenon, we are faced with a puzzle. The finding that the AAE is observed both in judgment measures and late eye-tracking measures (e.g. regression path and total times) suggests that the AAE in reading is driven by parsing failure or breakdown of some sort (Clifton & Staub, 2008). However, it is not clear that parallel or continuous explanations of the AAE

can be straightforwardly extended to acceptability judgment measures. In particular, it is not clear that differences in word predictability (Levy, 2008), statistical facilitation (Logacev & Vasishth, 2016a), or a time cost for computing a determinate attachment (Swets et al., 2008) would be severe enough to be reflected in judgment measures. Put simply, it appears that serial stochastic models explain why the ambiguity advantage effect looks like a reflection of parsing failure in reading (it is), but fail to explain why the effect is essentially graded in judgment measures. For underspecification or surprisal-based parsing models, the opposite is true: they can offer an explanation of the finding that both parses co-contribute to processing difficulty on a trial-to-trial basis, but the link between these models and acceptability measures is less clear.

To fully resolve this puzzle, a more complete parsing model is required; unfortunately, we are not in a position to offer such a model at present. Recall that parallel parsing models may be thought of in two distinct classes: those that posit that comprehenders maintain multiple, distinct symbolic representations of the input (Gibson, 1991; Gorrell, 1995; Hale, 2001; Levy, 2008), and those that can merge multiple distinct symbolic parses into a single representation that blends each parse (Cho, Goldrick & Smolensky, 2017; Christianson, Hollingworth, Halliwell, & Ferreira, 2001; Ferreira & Bailey, 2004; Rasmussen & Schuler, 2017; Tabor & Hutchins, 2004; van der Velder & De Kamps, 2006; Vosse & Kempen, 2000, 2009). At present, our data are in principle compatible with parallel models of each type.

To reconcile the eye-tracking and judgment data, parallel models that allow multiple traditional symbolic structures to be represented in parallel (Gibson, 1991; Gorrell, 1995; Hale, 2001; Levy, 2008) must adopt additional assumptions. For example,

it may be the case that the 'integration failure' that drives regressive eye-movements (Reichle, Warren & McConnell, 2009) may be triggered by an integration failure for any one of the representations under consideration by the parser.

However, if the parser is represents multiple syntactic analyses by blending them into a single representation, the eye-tracking data and the judgment data may be reconciled quite naturally. This is because these models essentially posit a composite representation of the input with two effective attachment sites of the relative clause, in a sort of 'blended' syntactic representation (Ferreira & Bailey, 2004; Ferreira & Patson, 2007; Lau & Ferreira, 2005; Slattery, Sturt, Christianson, Yoshida, & Ferreira, 2013). Given such a representation, parsing failure could arise in the unambiguous conditions for exactly the same reason it would arise in a serial and stochastic model: the resulting composite representation contains an ill-formed agreement dependency, which can both create integration failure (accounting for regressive eye-movements and re-reading in these conditions; Reichle et al., 2009) and the violation of a syntactic constraint (accounting for the effect in judgment measures). In this way, blend models can account for the AAE in reading time measures by maintaining that it reflects parse failure, as endorsed by the URM. However, they can account for the present results because they do not require that this failure reflect trial to trial variation in the attachment site selected by the comprehender.

A bit further afield, we note that a parser which operates by blending multiple syntactic representations by superimposing multiple attachments at once may offer additional benefits to the comprehender. For instance, we note that models that allow for syntactic blending bear some similarity to the 'overlay' mechanism that has been posited

by Ferreira and colleagues to explain how syntactic analyses that have been abandoned can persist and facilitate subsequent syntactic processing (Ferreira & Bailey, 2004; Lau, Ferreira & Bailey, 2007). The leading idea of this proposal is that a reanalyzed parse is not fully abandoned; instead, the parser 'overlays' the correct analysis on the existing structure, leading to a blended representation (see related discussion in Clifton & Staub, 2008; Ferreira & Patson, 2007; Slattery et al., 2013; Staub, 2007; Sturt, 2007). This mechanism accounts for the observation that the reanalyzed structure can be relatively easily reinstated and can otherwise continue to interfere in later processing (Christianson, Hollingworth, Halliwell, & Ferreira, 2001; Slattery et al., 2013; Staub, 2007; Sturt, 2007). This mechanism arguably already implies a limited sort of parallelism in the parser (Clifton & Staub, 2008), insofar as the parser can maintain blended representations of the input, allowing a single phrase to be associated with multiple attachment sites.

To be sure, there remain many open questions. Perhaps the most pressing question for future work concerns the conditions under which sentence ambiguity can facilitate processing and improve perceived acceptability. One interesting possibility is that the effect we see is limited to adjunct relations, such as the relative clause attachments we investigated. This possibility is suggested by evidence that the parser makes less determinate attachments for relative clauses than other types of syntactic dependency, perhaps owing to their status as adjunct or 'non-primary' syntactic relations (Frazier & Clifton, 1996; Swets et al., 2008). More generally, it is unclear what distinguishes the present contexts from cases where ambiguity seems to hinder, rather that facilitate, syntactic processing. Local coherence effects (Paape & Vasishth, 2016; Tabor, Galuntucci & Richardson, 2004) present one such a case. Local coherence presents as a

slowdown in reading times when readers consider a locally coherent, but globally incoherent, parse of the input. Put differently, local coherence effects can be described as contexts where perceived ambiguity creates a processing cost. Indeed, many different sorts of garden paths may be described in a similar fashion: in each case, ambiguity imposes a cost at the point when the sentence is disambiguated towards the dispreferred structure. More work is necessary to determine why ambiguity imposes a cost in these contexts, while improving the acceptability of relative clauses in the present study.

Conclusions

In this paper, we investigated the ambiguity advantage effect using a speeded acceptability judgment task. We offered a novel, signal-detection theoretic approach for investigating questions of syntactic representation created by the parser, and formal models for deriving distributions of underlying sentence acceptability given different representations of the input. We found that acceptability judgments in sentences with a relative clause attachment reflect the joint contribution of both possible analyses of that relative clause, a conclusion supported by both analyses of confidence ratings and reaction times, as well as mathematical modeling of our results. In all, our data and modeling suggest that the parser can represent multiple relative clause attachments simultaneously.

References

Abney, S. P. (1989). A computational model of human parsing. *Journal of Psycholinguistic Research*, *18*, 129-144.

Bader, M., & Häussler, J. (2010). Toward a model of grammaticality judgments1. *Journal of Linguistics*, *46*, 273-330.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*, 255–278.

Cho, P. W., Goldrick, M., Lewis, R. L., & Smolensky, P. (2017). Dynamic encoding of structural uncertainty in gradient symbols. To appear in *Proceedings of the 8th Workshop on Cognitive Modeling and Computational Linguistics* (CMCL 2018)

Christianson, K., Hollingworth, A., Halliwell, J. F., & Ferreira, F. (2001). Thematic roles assigned along the garden path linger. *Cognitive psychology*, *42*, 368-407.

Clifton Jr, C., & Frazier, L. (2010). When Are Downward-Entailing Contexts Identified? The Case of the Domain Widener Ever. *Linguistic Inquiry*, *41*, 681-689.

Clifton, C., & Staub, A. (2008). Parallelism and competition in syntactic ambiguity resolution. *Language and Linguistics Compass*, 2, 234-250.

DeCarlo, L. T. (2002). Signal detection theory with finite mixture distributions: Theoretical developments with applications to recognition memory. *Psychological Review*, *109*, 710-721.

Dillon, B., Staub, A., Levy, J., & Clifton Jr, C. (2017). Which noun phrases is the verb supposed to agree with?: Object agreement in American English. *Language*, *93*, 65-96.

Dube, C., & Rotello, C. M. (2012). Binary ROCs in perception and recognition memory are curved. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*, 130.

Earley, J. (1970). An efficient context-free parsing algorithm. *Communications of the ACM*, *13*, 94-102.

Egan, J., Schulman, A. I., & Greenberg, G. Z. (1959). Operating characteristics determined by binary decisions and by ratings. *The Journal of the Acoustical Society of America*, *31* 768-773.

Fanselow, G., & Frisch, S. (2006). Effects of processing difficulty on judgments of acceptability. In G. Fanselow, C. Féry, M. Schlesewsky, & R. Vogel (Eds.), *Gradience in grammar: Generative perspectives* (pp. 291–316). New York, NY: Oxford University Press.

Farmer, T. A., Cargill, S. A., Hindy, N. C., Dale, R., & Spivey, M. J. (2007). Tracking the continuity of language comprehension: Computer mouse trajectories suggest parallel syntactic processing. *Cognitive Science*, *31*, 889-909.

Ferreira, F., & Bailey, K. G. (2004). Disfluencies and human language comprehension. *Trends in cognitive sciences*, *8*(5), 231-237.

Ferreira, F., & Patson, N. D. (2007). The 'good enough' approach to language comprehension. *Language and Linguistics Compass*, *1*(1-2), 71-83.

Ferreira, F., & Henderson, J. M. (1991). Recovery from misanalyses of garden-path sentences. *Journal of Memory and Language*, *30*, 725-745.

Fodor, J. D., & Inoue, A. (1998). Attach anyway. In J.D. Fodor & F. Ferreira (eds.) *Reanalysis in sentence processing* (pp. 101-141). Springer: Dordrecht, Netherlands.

Frazier, L. (1979). *On comprehending sentences: Syntactic parsing strategies*. (Doctoral dissertation, University of Connecticut).

Frazier, L., & Fodor, J. D. (1978). The sausage machine: A new two-stage parsing model. *Cognition*, *6*, 291–325.

Frazier, L., & Clifton, C. (1996). *Construal*. Cambridge, MA: MIT Press.

Frazier, L., & Clifton Jr, C. (1998). Sentence reanalysis, and visibility. In J. Fodor & F. Ferreira (eds.), *Reanalysis in sentence processing* (pp. 143-176). Springer Netherlands.

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models.* Cambridge, United Kingdom: Cambridge University Press.

Gibson, E. (1991). *A computational theory of human linguistic processing: Memory limitations and processing breakdown* (Doctoral dissertation, Carnegie Mellon University). Retrieved from http://dl.acm.org/citation.cfm?id=124025

Gibson, E., & Pearlmutter, N. J. (2000). Distinguishing serial and parallel parsing. *Journal of Psycholinguistic Research*, *29*, 231-240.

Gorrell, P. (1995). *Syntax and parsing*. Cambridge: Cambridge University Press.

Grant M., Dillon, B., & Sloggett, S. (2014, September). *Ambiguity advantages in attachment and pronominal reference: Evidence from eye movements during reading*. Paper presented at the 20th annual Architectures and Mechanisms for Language Processing Conference, University of Edinburgh, United Kingdom.

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

Green, M. J., & Mitchell, D. C. (2006). Absence of real evidence against competition during syntactic ambiguity resolution. *Journal of Memory and Language*, *55*, 1-17.

Hale, J. (2001). A probabilistic Earley parser as a psycholinguistic model. In *Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies* (pp. 1-8). Association for Computational Linguistics.

Hammerly, C. M., Staub, A., & Dillon, B. (2018, April 25). The grammaticality asymmetry in agreement attraction reflects response bias: Experimental and modeling evidence. http://doi.org/10.17605/OSF.IO/6F34Y

Henderson, J. M., & Ferreira, F. (1993). Eye movement control during reading: Fixation measures reflect foveal but not parafoveal processing difficulty. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, *47*, 201.

Jurafsky, D. (1996). A probabilistic model of lexical and syntactic access and disambiguation. *Cognitive science*, *20*, 137-194.

Kimball, J. (1973). Seven principles of surface structure parsing in natural language. *Cognition*, *2*, 15–47.

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

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition, 106*, 1126–1177.

Lewis, R. L. (1998). Reanalysis and limited repair parsing: Leaping off the garden path. In J.D. Fodor & F. Ferreira (eds.) *Reanalysis in sentence processing* (pp. 247-285). Springer: Dordrecht, Netherlands.

Lewis, R. L. (2000). Falsifying serial and parallel parsing models: Empirical conundrums and an overlooked paradigm. *Journal of Psycholinguistic Research*, *29*, 241-248.

Lewis, R. L., & Vasishth, S. (2005). An activation-based model of sentence processing as skilled memory retrieval. *Cognitive Science, 29*, 375–419.

Logačev, P., & Vasishth, S. (2016a). A Multiple-Channel Model of Task-Dependent Ambiguity Resolution in Sentence Comprehension. *Cognitive Science*, *40*, 266-298.

Logačev, P., & Vasishth, S. (2016b). Understanding underspecification: A comparison of two computational implementations. *The quarterly journal of experimental psychology*, *69*, 996-1012.

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

Macmillan, N. A., & Creelman, C. D. (2005). *Detection theory: A user's guide*. Mahwah, NJ: Lawrence Earlbaum Associates.

Macmillan, N. A., & Kaplan, H. L. (1985). Detection theory analysis of group data: Estimating sensitivity from average hit and false-alarm rates. *Psychological Bulletin, 98*, 185-199

McElree, B. (2006). Accessing recent events. In B. Ross (Ed.), *Psychology of learning and motivation, vol. 46* (pp. 155–200). San Diego, CA: Elsevier Academic Press.

Paape, D., & Vasishth, S. (2016). Local coherence and preemptive digging-in effects in German. *Language and Speech*, *59*, 387-403.

Rasmussen, N. E., & Schuler, W. (2017). Left‐corner parsing with distributed Associative memory produces surprisal and locality effects. *Cognitive Science*.

Ratcliff, R., & McKoon, G. (2008). The diffusion decision model: theory and data for two-choice decision tasks. *Neural computation*, *20*, 873-922.

Ratcliff, R., Sheu, C. F., & Gronlund, S. D. (1992). Testing global memory models using ROC curves. *Psychological review*, *99*, 518-535.

Ratcliff, R., McKoon, G., & Tindall, M. (1994). Empirical generality of data from recognition memory receiver-operating characteristic functions and implications for the global memory models. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*, 763.

Reichle, E. D., Warren, T., & McConnell, K. (2009). Using EZ Reader to model the effects of higher level language processing on eye movements during reading. *Psychonomic bulletin & review*, *16*, 1-21.

Schütze, C. T., & Sprouse, J. (2014). Judgment data. In R. Podesva and D. Sharma (eds.), *Research methods in linguistics* (pp. 27-50). Cambridge University Press: Cambridge.

Slattery, T. J., Sturt, P., Christianson, K., Yoshida, M., & Ferreira, F. (2013). Lingering misinterpretations of garden path sentences arise from competing syntactic representations. *Journal of Memory and Language*, *69*, 104-120.

Sprouse, J. (2008). The differential sensitivity of acceptability judgments to processing effects. *Linguistic Inquiry, 39*, 686–694.

Staub, A. (2007). The return of the repressed: Abandoned parses facilitate syntactic reanalysis. *Journal of memory and language*, *57*, 299-323.

Stewart, A. J., Holler, J., & Kidd, E. (2007). Shallow processing of ambiguous pronouns: Evidence for delay. *The Quarterly Journal of Experimental Psychology*, *60*, 1680-1696.

Sturt, P. (2007). Semantic re-interpretation and garden path recovery. *Cognition*, *105*(2), 477-488.

Swets, B., Desmet, T., Clifton, C., & Ferreira, F. (2008). Underspecification of syntactic ambiguities: Evidence from self-paced reading. *Memory & Cognition*, *36*, 201-216.

Tabor, W., & Hutchins, S. (2004). Evidence for self-organized sentence processing: Digging-in effects. *Journal of Experimental Psychology-Learning Memory and Cognition*, *30*, 431-449.

Tabor, W., Galantucci, B., & Richardson, D. (2004). Effects of merely local syntactic coherence on sentence processing. *Journal of Memory and Language*, *50*(4), 355-370.

Traxler, M. J., Pickering, M. J., & Clifton, C. (1998). Adjunct attachment is not a form of lexical ambiguity resolution. *Journal of Memory and Language*, *39*, 558-592.

Trueswell, J. C., Tanenhaus, M. K., & Garnsey, S. M. (1994). Semantic influences on parsing: Use of thematic role information in syntactic ambiguity resolution. *Journal of memory and language*, *33*, 285-318.

Van der Velde, F., & De Kamps, M. (2006). Neural blackboard architectures of combinatorial structures in cognition. *Behavioral and Brain Sciences*, *29*, 37-70.

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

Van Gompel, R. P. G., Pickering, M. J., & Traxler, M. J. (2000). Unrestricted race: A new model of syntactic ambiguity resolution. In A. Kennedy, R. Radach, D. Heller, & J. Pynte (Eds.), *Reading as a perceptual process* (pp. 621–648). Oxford, England: Elsevier.

Van Gompel, R. P., Pickering, M. J., & Traxler, M. J. (2001). Reanalysis in sentence processing: Evidence against current constraint-based and two-stage models. *Journal of Memory and Language*, *45*, 225-258.

Van Gompel, R. P., Pickering, M. J., Pearson, J., & Liversedge, S. P. (2005). Evidence against competition during syntactic ambiguity resolution. *Journal of Memory and Language*, *52*, 284-307.

Vasishth, S., Nicenboim, B., Chopin, N., & Ryder, R. (submitted). Bayesian hierarchical finite mixture models of reading times: A case study

Vasishth, S., Bruessow, S., Lewis, R. L., & Drenhaus, H. (2008). Processing polarity: How the ungrammatical intrudes on the grammatical. *Cognitive Science*, *32*, 685-712.

Vosse, T., & Kempen, G. (2000). Syntactic structure assembly in human parsing: a computational model based on competitive inhibition and a lexicalist grammar. *Cognition*, *75*, 105-143.

Vosse, T., & Kempen, G. (2009). In defense of competition during syntactic ambiguity resolution. *Journal of psycholinguistic research*, *38*, 1.

Wagenmakers, E. J., Ratcliff, R., Gomez, P., & Iverson, G. J. (2004). Assessing model mimicry using the parametric bootstrap. *Journal of Mathematical Psychology*, *48*, 28-50.

Warner, J., & Glass, A. L. (1987). Context and distance-to-disambiguation effects in ambiguity resolution: Evidence from grammaticality judgments of garden path sentences. *Journal of Memory and Language*, *26*, 714-738.

Weinberg, A. (1993). Parameters in the theory of sentence processing: Minimal commitment theory goes east. *Journal of Psycholinguistic Research*, *22*, 339-364.

Winograd, T. (1983). *Language as a cognitive process, vol. 1: Syntax.* Reading, MA: Addison Wesley.

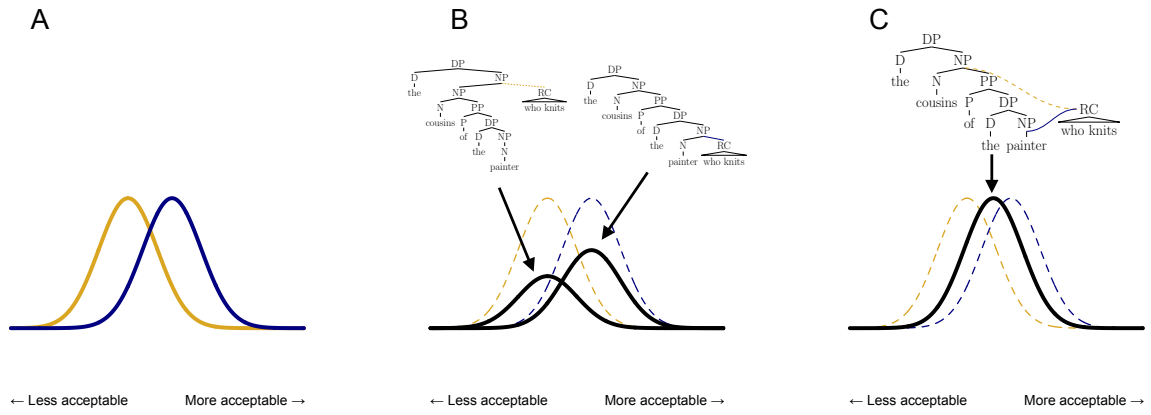

Figure 1: Visual summary of *Acceptability* distributions. Panel A: *Acceptability* distributions for Pure sentences. Yellow solid line (left) is hypothesized distribution of *Acceptability* for pure ungrammatical (NOMATCH), blue solid line (right) is distribution for pure grammatical sentences (MULTIMATCH). Panel B: *Acceptability* distribution (solid black lines) for Mixed sentences under a serial parsing model, reflecting the distribution of *Acceptability* on different types of trials. Dashed lines represent the same distributions as in Panel A. Panel C: *Acceptability* distribution (solid black lines) for Mixed sentences under a parallel parsing model, reflecting a weighted sum of samples from pure distributions. Dashed lines represent the same distributions as in Panel A.

Figure 2: Signal detection model for a simple task (left panel); Receiver operating characteristic curve (middle panel); zROC (right panel).

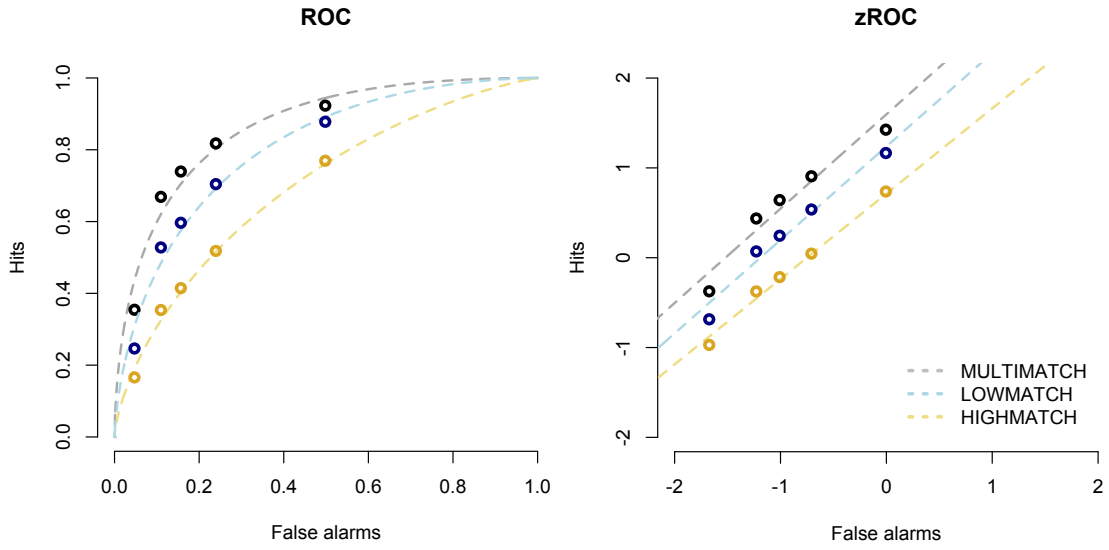

Figure 3: Empirical ROC and *z*ROC functions for all three empirical comparisons. Open circles plot observed hit/false alarm rates at each level of confidence. The hit rates (*y*-coordinates) come from MULTIMATCH, HIGHMATCH, and LOWMATCH conditions. In all cases, the false alarm rate (*x*-coordinates) come from NOMATCH conditions. Dashed lines indicate best-fit UVSDT model to all four conditions. Grey (topmost line) indicates MULTIMATCH comparison; blue indicates LOWMATCH comparison; yellow (bottommost line) indicates HIGHMATCH comparison.

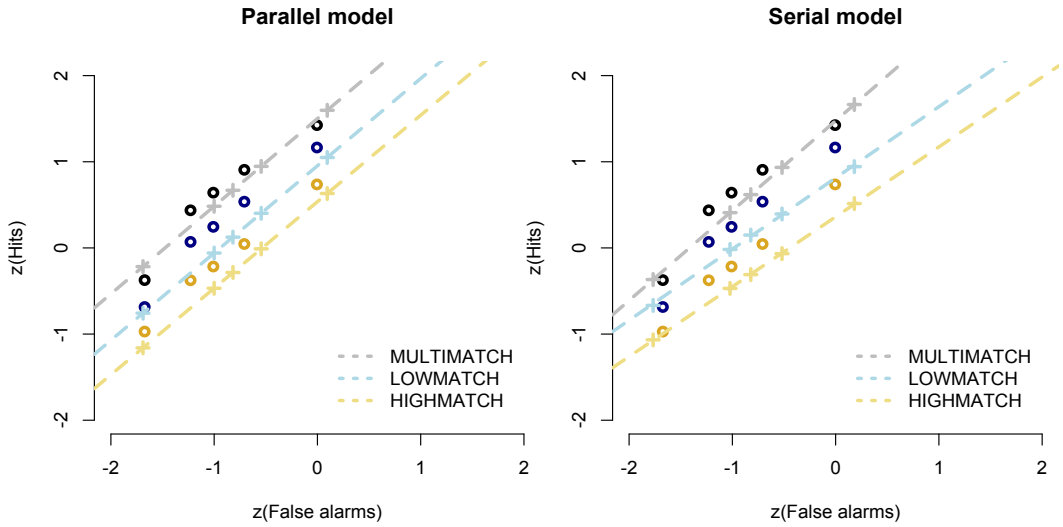

Figure 4: zROC comparison of observed data (open circles) to model predictions (crosses and dashed lines). Best-fit parallel model is plotted in the left panel; best-fit serial model on the right. Grey (topmost line) indicates MULTIMATCH comparison; blue indicates LOWMATCH comparison; yellow (bottommost line) indicates HIGHMATCH comparison.

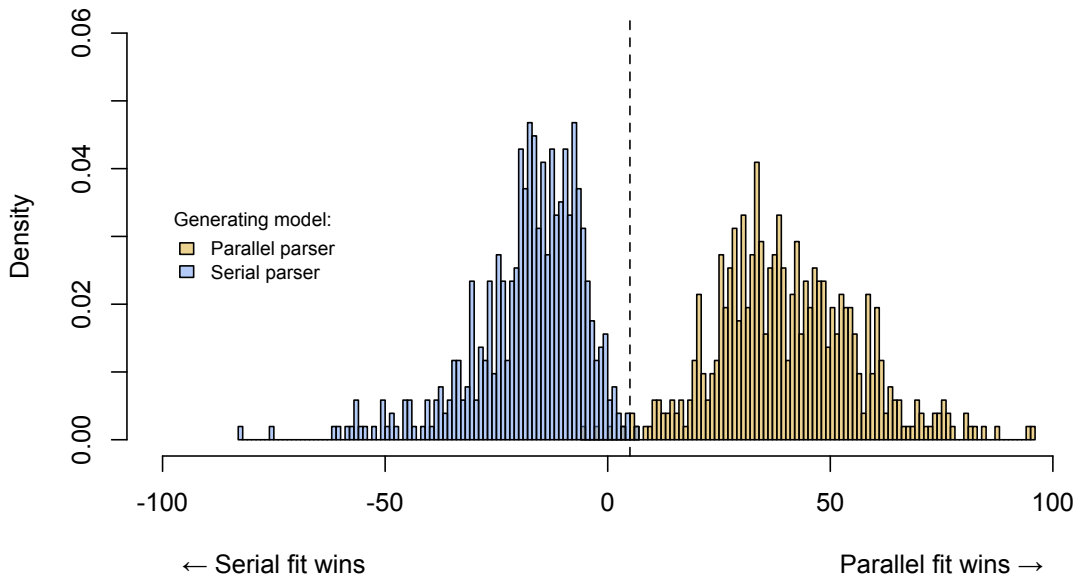

Figure 5: **Histogram of bootstrapped differences in $-2\mathcal{L}(\theta)$, serial fit minus parallel fit**. Negative values indicate that the serial model provided a better fit to a simulated data set; positive values indicate that the parallel model was a better fit. Vertical dashed line indicates optimal $-2\mathcal{L}(\theta)$ difference for deciding between models.

Condition	% Accepted	"Yes" responses		"No" responses	
		RT	Confidence	RT	Confidence
NOMATCH	0.16	1531 (87)	2.0 (0.10)	1185 (35)	2.5 (0.04)
MULTIMATCH	0.74	1267 (34)	2.4 (0.04)	1458 (59)	1.9 (0.07)
LOWMATCH	0.60	1356 (41)	2.3 (0.04)	1464 (52)	2.0 (0.06)
HIGHMATCH	0.41	1455 (49)	2.2 (0.05)	1352 (41)	2.2 (0.05)

Table 1: **Ambiguous Conditions are Judged More Quickly and More Confidently.** Summary of % accepted, reaction time (RT) and confidence ratings for 'yes' and 'no responses to all conditions. Boxed cells indicate correct response. Confidence ranged from 1 (not at all confident) to 3 (very confident). By-participant standard error in parentheses.

	% Yes		Confidence		RT	
	β		β		β	
Amb	.81 (±.24)	3.4	-.09 (±.11)	-.9	.08 (±.04)	2.0
Height	-.91 (±.20)	-4.4	.13 (±.08)	1.6	.00 (±.02)	0.2
Gram	-3.37 (±.25)	-13.4	.14 (±.09)	1.6	-.03 (±.03)	-0.8
Correct	-	-	.55 (±.07)	7.8	-.07 (±.02)	-3.4
Amb × Acc	-	-	-1.15 (±.27)	-4.2	.34 (±.09)	3.7
Height × Acc	-	-	-.47 (±.15)	-3.2	.17 (±.05)	3.6
Gram × Acc	-	-	.25 (±.22)	1.1	-.09 (±.08)	-1.2

Table 2: Summary of mixed-effect regression analyses for all three dependent measures in the judgment task. Shaded cells represent reliable effects at $t/Z = 2$.

	Accuracy: d_a	Slope: 1/s
MULTIMATCH	1.52	1.05
LOWMATCH	1.19	1.04
HIGHMATCH	.75	.95
MULTIMATCH *v.* LOWMATCH	.33 [.23-.44]	.01 [-.08-.12]
MULTIMATCH *v.* HIGHMATCH	.78 [.61-.95]	.09 [.00-.19]
LOWMATCH *v.* HIGHMATCH	.45 [.28-.62]	.08 [-.02-.19]

Table 3: Results of UVSDT analysis, including accuracy (measured in d_a.) and slope (1/s) by comparison. The lower three rows represent the mean difference in estimated parameters between ROCs along with a 95% confidence interval on the difference across bootstrap samples. Shaded cells indicate a significant difference at $\alpha = 0.05$.