Models of retrieval in sentence comprehension: A computational evaluation using Bayesian hierarchical modeling

Bruno Nicenboim
Department of Linguistics, University of Potsdam,
Potsdam, Germany
e-mail:bruno.nicenboim@uni-potsdam.de

Shravan Vasishth
Department of Linguistics, University of Potsdam, Potsdam, Germany
CEREMADE (Centre de Recherche en Mathématiques de la Décision), Université Paris-Dauphine, Paris, France

Draft of October 22, 2016.

Author Note

We thank Hedderik van Rijn for the useful discussion during Groningen Spring School on Cognitive Modeling, 2016, and Pavel Logačev for insightful comments on the manuscript. This work was supported by Potsdam Graduate School, and the University of Potsdam.

Abstract

Research on similarity-based interference has provided extensive evidence that the formation of dependencies between non-adjacent words relies on a cue-based retrieval mechanism. There are two different models that can account for one of the main predictions of interference, i.e., a slowdown at a retrieval site, when several items share a feature associated with a retrieval cue: Lewis and Vasishth's (2005) activation-based model and McElree's (2000) direct access model. Even though these two models have been used almost interchangeably, they are based on different assumptions and predict differences in the relationship between reading times and response accuracy. The activation-based model follows the assumptions of the ACT-R framework, and its retrieval process behaves as a lognormal race between accumulators of evidence with a single variance. Under this model, accuracy of the retrieval is determined by the winner of the race and retrieval time by its rate of accumulation. In contrast, the direct access model assumes a model of memory where only the probability of retrieval can be affected, while the retrieval time is constant; in this model, differences in latencies are a by-product of the possibility of backtracking and repairing incorrect retrievals. We implemented both models in a Bayesian hierarchical framework in order to evaluate them and compare them. We show that some aspects of the data are better fit under the direct access model than under the activation-based model. We suggest that this finding does not rule out the possibility that retrieval may be behaving as a race model with assumptions that follow less closely the ones from the ACT-R framework. We show that by introducing a modification of the activation model, i.e, by assuming that the accumulation of evidence for retrieval of incorrect items is not only slower but noisier (i.e., different variances for the correct and incorrect items), the model can provide a fit as good as the one of the direct access model.

Keywords: cognitive modeling; sentence processing; working memory; cue-based retrieval; similarity-based interference; Bayesian hierarchical modeling

Models of retrieval in sentence comprehension: A computational evaluation using Bayesian hierarchical modeling

There is strong evidence that the formation of syntactic dependencies between non-adjacent words relies on the memory system. An example is the so-called locality effect: increasing the distance between co-dependents (such as subjects and verbs) tends to lead to greater processing difficulty (Gibson, 2000; Grodner & Gibson, 2005). Research on interference makes a similar point: the speed and/or accuracy of dependency completion can be adversely affected by the presence of items in memory that are similar to the retrieval target (among others: Gordon, Hendrick, & Levine, 2002; Van Dyke & Lewis, 2003; Van Dyke, 2007; Jäger, Engelmann, & Vasishth, 2015; Nicenboim, Engelmann, Suckow, & Vasishth, submitted; Vasishth, Brüssow, Lewis, & Drenhaus, 2008). Such a central role for memory in sentence comprehension is well-motivated: it is implausible that the parser could keep track of a large and in principle unbounded inventory of the dependencies that can be found in a sentence, since they easily exceed the amount of information that can be held in the focus of attention (McElree & Dosher, 1989; McElree, 2006; Cowan, 1995; Oberauer, 2013; Marcus, 2013). The evidence from studies investigating similarity-based interference (see the meta-analysis of published studies in Jäger, Engelmann, & Vasishth, submitted) strongly suggests that dependency completion relies on a *content-addressable cue-based retrieval mechanism* that is subject to interference (McElree, 2000; Van Dyke & Lewis, 2003; Lewis, Vasishth, & Van Dyke, 2006). Similarity-based interference is a phenomenon that is not unique to language, and occurs when several items share a feature associated with a retrieval cue. A major implication is that the retrieval mechanism employed for the creation of linguistic dependencies is similar to the one utilized in non-language domains.

There are multiple implementations compatible with such a content-addressable cue-based retrieval mechanism in sentence processing. As a verbally stated model, this type of mechanism would entail that when retrieval cues fully match the target of retrieval,

similarity-based interference would cause an inhibitory effect, that is, an increase of processing difficulty at the retrieval of a dependent. This processing difficulty would be reflected in longer reading times and lower accuracy. However, in some cases, shorter reading times have been observed when increased processing difficulty was clearly expected (Van Dyke & McElree, 2006; Nicenboim, Vasishth, Gattei, Sigman, & Kliegl, 2015; Nicenboim, Logačev, Gattei, & Vasishth, 2016). In these cases, it is usually assumed that the fast reading times are a consequence of a shallow parse (due to, for example, good-enough processing, Ferreira, Bailey, & Ferraro, 2002) caused by cognitive overload. There can be good reasons to assume that shorter reading times are associated with increased difficulty, for example, when shorter reading times co-occur with lower comprehension accuracy (Van Dyke & McElree, 2006) or lower working memory capacity (Nicenboim et al., 2015; Nicenboim et al., 2016). However, the trade-off between reading times and comprehension accuracy is usually left underspecified.

There are two models that make explicit the relationship between reading times and retrieval accuracy, and even though they are sometimes not differentiated, they implement the content-addressable cue-based retrieval mechanism in a different way: Lewis and Vasishth's (2005) model, which we will refer to as the *activation-based* model, and McElree's (2000) model, which we will refer to as the *direct access* model. While the models are similar, as we explain in detail later, they have different implications for retrieval processes in sentence comprehension. The activation-based model assumes a process that resembles a race model (Audley & Pike, 1965; Vickers, 1970), where evidence for each retrieval candidate is accumulated with different rates. This race determines both the latencies and the retrieval accuracy. By contrast, the direct access model assumes that retrieval candidates have different levels of *availability*, which is the probability that a memory representation is retained. Availability determines only the accuracy of the retrieval and not the latency. In this model, a difference in latency between two conditions is a by-product of a mixture of directly accessed items, and retrievals that were initially

incorrect, but they are reanalyzed leading to a correct retrieval.

The goal of this paper is to unpack the quantitative predictions of the activation-based and direct access models by implementing them in a Bayesian hierarchical framework. This will allow us to compare their relative fit to a representative dataset and to assess their validity as models of retrieval that can account for similarity-based interference. We used the data from Nicenboim et al. (submitted), which investigated similarity-based interference from the number feature using two relatively high-powered self-paced reading experiments. The data in this study include two dependent measures: (i) reading times for the critical region where retrieval from memory is assumed to occur, and (ii) accuracies in a comprehension task that targets specific dependency relations through a multiple choice task. This dataset is especially suitable for our modeling purposes because, apart from Van Dyke (2007, who also evaluated some of the dependencies), this is the only dataset that we are aware of that uses comprehension questions to directly assess the resolution of the dependencies. As we explain in detail later, these two dependent measures (reading times and accuracy) are necessary for evaluating the models.

Nicenboim et al. (submitted) used stimuli like (1). There were two conditions, high vs. low interference, which were assumed to affect the dependency between the subject (i.e., *Der Wohltäter* "The philanthropist") and the verb (i.e., *begrüßt hatte* "had greeted"). In the high interference condition, two nouns intervened between these two co-dependents that had the same number marking as the target noun, the subject of the sentence, namely, singular marking. In the low interference case, the two intervening nouns had plural marking while the target noun remained singular. In German, the verb (i.e.,*begrüßt hatte*) agrees in number with its subject; in the high interference condition, the retrieval cue set at the verb to seek out a singular noun would match three nouns. By contrast, in the low interference condition, only one noun matches this retrieval cue. Thus, reading time at the critical region, the verb *begrüßt hatte*, provides an estimate of any interference effect.

Each target sentence was followed by a question that queried either the subject of the

matrix verb (e.g., “sat”), the subject of the embedded verb (e.g.,“had greeted”), or the object of the embedded verb. The possible answers were provided in multiple-choice format in pseudo-randomized order. For all the questions, participants had the option to answer “I don’t know”, when they did not remember or could not answer.

(1) a. High Interference

Der Wohltäter, der den Assistenten
The.sg.nom philanthropist, who.sg.nom the.sg.acc assistant (of)
des Direktors **begrüßt hatte**, saß später im
the.sg.gen director **greeted had.sg**, sat.sg later in the
Spendenausschuss.
donations committee.

‘The philanthropist, who had greeted the assistant of the director, sat later in the donations committee.’

b. Low Interference

Der Wohltäter, der die Assistenten
The.sg.nom philanthropist, who.sg.nom the.pl.acc assistants (of)
der Direktoren **begrüßt hatte**, saß später im
the.pl.gen directors **greeted had.sg**, sat.sg later in the
Spendenausschuss.
donations committee.

‘The philanthropist, who had greeted the assistants of the directors, sat later in the donations committee.’

Nicenboim et al. (submitted) found an inhibitory effect of similarity-based interference for the retrieval of the subject (“the philanthropist”) when it shared the number feature singular with other competitor NPs (“the assistants”, “the directors”); this suggests that retrieval in high interference conditions took longer than in low interference conditions. Given that the auxiliary verb (*hatte* “had”) is morphologically marked as singular, the longer reading times at the auxiliary verb is consistent with cue-based retrieval.

Both the activation-based and the direct access models make the correct predictions regarding the average behavior—both predict an inhibitory interference effect. However,

these two accounts differ in the way that correct and incorrect retrievals occur, and these different underlying mechanisms can be investigated using the data from Nicenboim et al. (submitted).

We first describe the models qualitatively, but in order to unpack the predictions of these two models we later provide a more formal presentation. We then evaluate the models quantitatively by examining the relationship between reading times and accuracy.

There are two main findings in the present study: First, the direct access model provides a better fit to the data than the activation-based model corresponding to Lewis & Vasishth, 2005. Second, we show that a variation of the activation-based model fits the data as well as the direct access model, and also provides a reasonable model of the underlying generative process.

Overview of the activation-based and direct access model

The activation-based model as implemented by Lewis and Vasishth (2005) is a computational model of sentence processing in which dependencies of non-adjacent elements are created via a content-addressable cue-based retrieval mechanism. This model was realized in ACT-R (Anderson et al., 2004), which is a general cognitive architecture used to model a vast variety of cognitive phenomena. This means that sentence processing depends on the application of general cognitive principles to the specialized task of sentence parsing. Being a computational model, it provides quantitative predictions of retrieval speed and accuracy.

The predictions regarding interference, locality, and some antilocality effects of Lewis and Vasishth's (2005) original model have been investigated using simulation (Lewis & Vasishth, 2005; Vasishth et al., 2008). In addition, simplified versions of the model, which focused on certain aspects of the architecture and evaluated some of the assumptions of the original model, have also been used (e.g., Dillon, Mishler, Sloggett, & Phillips, 2013; Dillon et al., 2014; Kush & Phillips, 2014; Jäger et al., 2015; Vasishth & Lewis, 2006;

Nicenboim et al., 2016; Engelmann, 2015; Parker & Phillips, 2016).

Crucially, the activation-based model provides an account of the relationship between reading times at the dependency resolution site and the accuracy of the resolution. This is so because dependency creation relies on the retrieval of the correct item from memory; in ACT-R terms, what is retrieved is a chunk. The chunk with the highest activation is retrieved and its activation level determines the retrieval time. We provide next an informal explanation of the key aspects of our implementation of the activation-based model. We do this using example (1) from Nicenboim et al. (submitted). We show that the activation-based model can explain similarity-based interference effects, predicting inhibitory interference (i.e., an increase in processing difficulty) when a competitor NP matches the singular number feature of the target of retrieval.

The main assumptions of the model are that (i) words are encoded in memory as bundles of features (as in Nairne, 1990; Oberauer & Kliegl, 2006) that include lexical, semantic, and syntactic information, and that (ii) retrieval cues are used to identify the "correct" chunk from memory: If retrieval cues (which are feature specifications) match with the features of a chunk in memory, the chunk gets a boost in activation, and if cues mismatch a chunk's features, activation is decreased. Such a mechanism would always retrieve the correct item. However, due to random noise in the system, activation fluctuates randomly from trial to trial, so that despite a cue match with a target, a competitor could have higher activation and could end up being retrieved. An alternative possibility is that all candidate chunks in memory fall below a retrieval activation threshold (a parameter in ACT-R); in this case, retrieval would fail.

As an example, consider the auxiliary verb (*hatte*, "had.sg") of (1). This is the region where an interference effect was seen in Nicenboim et al. (submitted). In both sentences (1a) and (1b), there is a dependency between this verb and its subject ("the philanthropist"), and the only difference between the sentences is that the intervening NPs ("the assistant/s", "the director/s") appear in singular in (1a) and in plural in (1b). The

activation-based model assumes that the feature information of each item such as category, case, number, gender, and so forth is encoded in memory. When the embedded verb (*begrüßt hatte*, "greeted had.sg") is being read, an attempt is made to retrieve the subject. The verb provides cues such as *NP*, *nominative* (notice that case is encoded in the determiner of the NP in German), *singular*, among others features required from the target of the retrieval. For each cue, a limited amount of activation (called the maximum associated strength or MAS) is spread among the target and the competitors that are stored in memory. The MAS determines the strengths of association from each cue to each item in memory. This strength of association represents how uniquely the cue identifies a target. This means that in the low interference condition (1b), the strength of association of the singular cue with the target is determined by the maximum activation associated with this cue (since the cue fully identifies the item). In the high interference condition (1a), however, the target ("the philanthropist") and the competitors ("the assistant", "the director") will be assigned some smaller part of the maximum activation (this is the so-called fan effect, for details, see Anderson & Reder, 1999), and thus their strength of association of *singular* will be smaller than the maximum activation. This is regardless of the fact that in both conditions, there is another cue that uniquely identifies the target, namely, being nominative: In both conditions the target also receives activation due to the strength of association with the cue nominative. As it is shown more formally in the next section, this means that the target would receive (on average) more total spreading activation than the competitors. See Figure 2 for a schematic that explains this.

One crucial aspect of the activation-based model is that retrieval is not deterministic because it depends on activation that fluctuates from trial to trial due to noise. Given that both latency and probability are affected by activation, we will show later that the retrieval process is similar to a race of accumulators (among many others: Audley & Pike, 1965; Vickers, 1970; Usher & McClelland, 2001): Each item in memory is assigned an *accumulator of evidence* for its retrieval, where the activation of each item acts as the rate

of accumulation. The accumulator that reaches the threshold of evidence first determines which item is retrieved and with which latency. Furthermore, some of the assumptions of ACT-R can allow us to frame the retrieval process as one of simplest accumulator models: the lognormal race model with a single variance for the noise associated with target, competitors, and failure accumulators (Heathcote & Love, 2012; Rouder, Province, Morey, Gomez, & Heathcote, 2014). A variant that we will consider later is a model with two separate variances, one for the target accumulator, and one for the competitors and failure accumulators.

It should be noted that a content-addressable system does not necessarily entail a race between items in memory, and there are other models that are also compatible with a content-addressable cue-based retrieval mechanism. The cue-based retrieval model proposed in Van Dyke and McElree (2006) is based on McElree and colleagues' previous work (e.g., McElree, 2000; McElree, Foraker, & Dyer, 2003) and, while it does not assume a race model, it shares with the activation-based model some of the assumptions of the cue-based retrieval mechanism: Words are also encoded in memory as feature bundles, and retrieval cues are used to distinguish the target from the competitors. In addition, even though the mechanism is different from the one in the activation-based model, here too the probability of retrieving a particular item from memory given the retrieval cues is a function of the degree of the match between the cues and the item, reduced by the degree to which the cues match other competitor items in memory. However, in contrast with the activation-based account, cues are supposed to enable *direct access* to relevant memory representations. This means not only that there is no serial search between items in memory, but that the distribution of access time is independent of the degree of match between item and cue, and regardless of the quality or strength of the representation of the item in memory (McElree, 2000).

It is not uncommon, however, that both poorer accuracy and longer reading times associated with similarity-based interference are taken as evidence for McElree's (2000)

direct access model as well as for Lewis and Vasishth's (2005) activation-based model. For example Van Dyke and McElree (2006) write:

> The current experiment supports a retrieval-based account of interference effects in sentence processing, one that is compatible with the hypothesis that a cue-based retrieval mechanism mediates the creation of grammatical dependencies during parsing. One such mechanism has been proposed in Van Dyke and Lewis (2003; see also Lewis and Vasishth, 2005; Van Dyke, 2002), in which parsing success depends on the extent to which required constituents can be retrieved from working memory. On this account grammatical heads provide retrieval cues that are used to access previously stored items via a content-addressable retrieval process (McElree, 2000, 2006; McElree, Foraker, & Dyer, 2003).

A slowdown in self-paced reading or eyetracking-while-reading can also be taken as evidence for direct access, since processing speed may be affected by differences in the likelihood of recovering an item from memory (McElree, 1993; McElree et al., 2003). This is because McElree (1993) assumes that after a misretrieval, that is, an incorrect or failed retrieval, the parser can often backtrack to reprocess the retrieval and reach the appropriate analysis. This would mean that a correct interpretation of a dependency could be arrived at because the correct dependent was retrieved at the first attempt or, alternatively, because a wrong dependent was retrieved initially but the parser backtracked and retrieved the correct one. Given that backtracking should take some additional time, latencies associated with the correct responses would be a mixture of fast directly accessed dependents and retrievals slowed down due to the time needed for backtracking. Since interference adversely affects retrieval probabilities, the proportion of errors would be higher in high interference conditions. This would entail a higher proportion of backtracking and hence slower latencies in the mixture of correct responses and on average,

high interference conditions would show longer reading times than low interference conditions.

Since the assumed constant distribution of retrieval times may not be observable in reading for comprehension, evidence compatible with the direct access model but incompatible with the activation-based model comes only from findings of speed-accuracy trade-off (SAT) procedure on rapid grammaticality judgment task (e.g. McElree, 2000; McElree et al., 2003; Van Dyke & McElree, 2011). In this task, participants need to judge a sentence as either grammatical or ungrammatical, and their judgment process is interrupted with a cue to respond (typically a tone) after varying amounts of time (Reed, 1973; Wickelgren, 1977; and see also: Foraker & McElree, 2011). However, this is a meta-linguistic task and it would be desirable if independent support for the direct access model could be found with a reading-for-comprehension task. In addition, the conclusion that retrieval time is constant requires arguing for a null result in one of the parameters of the SAT model. For SAT procedures, accuracy is modeled as a function of three parameters corresponding to the three phases of the SAT curve: (i) the asymptotic level of performance, (ii) the intercept or the point in time where performance is different from chance, and (iii) the rate at which accuracy grows from chance to asymptote. While the presence of the effect in the asymptote, so that an increase on interference lowers the asymptote, is evidence for the reduction of the probability of accessing the target, the lack of evidence for changes in the rate and intercept must be taken as evidence for no effect on the speed of the retrieval. It could be, however, that the differences were too small to be detected.

Since both models give virtually identical predictions for non-SAT (self-paced reading or eye-tracking) experiments for averages, slowdowns product of similarity-based interference have been taken as evidence for the two models. However, the two models assume different relationships between retrieval times and responses. It is important to assess the fit to the data of each model, since each one is compatible with different memory

retrieval mechanisms. The activation-based model uses the declarative retrieval module of ACT-R, which has been shown to be able to account for many memory-related phenomena (e.g., Anderson, Bothell, Lebiere, & Matessa, 1998; Anderson & Reder, 1999; Van Rijn & Anderson, 2003). In addition, the characteristics of the retrieval process allow us to model it as a race of accumulators (Vickers, 1970), which places the model under a sequential sampling framework (such as the drift diffusion model: Ratcliff, 1978; Ratcliff & McKoon, 2008; the leaky competitive accumulator: Usher & McClelland, 2001; linear deterministic models: Brown & Heathcote, 2008, among others; Heathcote & Love, 2012). In the sequential sampling framework, decisions (such as which is the right dependent that needs to be retrieved) are considered a process of noisy accumulation of evidence.

In contrast to the activation-based model, the direct access model assumes a bipartite architecture for retrieval (e.g., McElree & Dosher, 1989; McElree, 2006): Items within focal attention are accessed quickly, but all other items outside attention are accessed more slowly and with the same retrieval speed. The direct access model allows items that are outside the focus of attention to vary only in their level of availability, i.e. their probability of retrieval.

Implementation of the models

In order to distinguish between the activation-based and direct access models, we implemented them as hierarchical Bayesian models in Stan (Stan Development Team, 2016b).[1] See Lee (2011) for a similar approach to cognitive modeling. Implementing these models affords several advantages: (i) we can investigate (through posterior predictive checking, explained below) whether the data could have been generated by the models; (ii) we can determine how each model's parameters were affected by interference effects; and

[1] Given the existence of models that assume that some aspect of the mind is Bayesian, such as Bayesian approaches to parsing (such as Kleinschmidt, Fine, & Jaeger, 2012; Traxler, 2014; Myslín & Levy, 2016) or to word learning (see, for example Xu & Tenenbaum, 2007), it is important to note that we are using Bayesian methods as a flexible and interpretable way of extending models of cognitive processes (Lee, 2011; Shiffrin, Lee, Kim, & Wagenmakers, 2008). This approach is orthogonal to the question of whether the mind does or does not do Bayesian inference.

(iii) the quality of fit of the two models can be compared using cross-validation.

The benefits of using hierarchical Bayesian modeling are two-fold: (i) The Bayesian aspects of the models mean that the model incorporates the general advantages of Bayesian inference, such as the use of credible intervals instead of confidence intervals, and the possibility of fitting complex non-linear models (see Nicenboim & Vasishth, 2016, for an extended discussion), and (ii) the hierarchical aspects entail that the models take both between- and within-group variances into account and pool information via shrinkage (Gelman, Hill, & Yajima, 2012). This means that we avoid overfitting the data and at the same time we avoid averaging and losing valuable information about group-level variability (Gelman & Hill, 2007). In the next section we provide a more formal presentation of the assumptions and details of the implementation of both models than the one given in the introductory overview.

The activation-based model

Lewis and Vasishth's (2005) activation-based account is based on ACT-R (Anderson et al., 2004) which is explicit about the effect of similarity-based interference in both latency and accuracy. Even though ACT-R is a general cognitive architecture, we focus only on the ability of the architecture to explain the retrieval process under interference and not on the full framework. A detailed description of the full ACT-R framework is provided in Anderson et al. (2004) and Anderson and Lebiere (1998).

In ACT-R, all chunks (which in our case are words or phrases) in memory have an activation level that represents their strength in memory. In the classical ACT-R framework (as opposed to RACE/A extension proposed by van Maanen, van Rijn, & Taatgen, 2011), the activation of the chunk in memory is calculated at the onset of a particular retrieval request. The total activation, A_c, assigned to each chunk c stored in memory is the sum of the base-level activation, B_c, which describes the history of usage of a chunk; the spreading activation, S_c, which represents the influence of the cues in identifying

the chunk; a penalty component, P_c, which is the activation deducted in case of mismatch; and a random noise component, ϵ, that makes the retrieval a probabilistic process:

$$A_c = B_c + S_c + P_c + \epsilon \tag{1}$$

Similarity-based interference affects the value of the spreading activation, S_c. This value, S_c, represents the sum of each of the strengths of association, $S_{c,u}$, from each cue u to chunk c, assuming N different cues, weighted by the importance of the cue, W_u; see Equation (2). The value of each $S_{c,u}$ will range between zero, meaning that cue u is not helpful for distinguishing the target between competitors in memory (e.g., being a word is certainly a feature required, but it provides no useful association to the target), to the maximum associated strength (MAS), which represents the gain in activation when the cue unequivocally distinguishes the target (since the cue matches a feature that no competitor possesses); see Figure 1.[2]

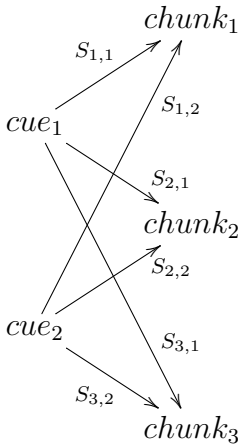

Figure 1. Graph showing the associations between two cues and three chunks.

[2]The value of the strengths of association, $S_{c,u}$, is defined in Anderson and Reder (1999) as depending on the probability of needing chunk c when cue u is present, $P(c|u)$, so that $S_{c,u} = MAS + log(P(c|u))$. The problem is that outside experiments where the association between cue and chunk (so-called fan effect) is highly controlled, $P(c|u)$ is very difficult to estimate. In addition, since $P(c|u) < 1$, it follows that $log(P(c|u)) < 0$, this means that MAS needs to be set to avoid a negative value of spreading activation. Since we do not need to know the exact value of S_c, for our purposes, distinguishing $S_{c,u} < MAS$ from $S_{c,u} = MAS$ is enough.

$$S_c = \sum_{u=1}^{N} W_u \cdot S_{c,u} \tag{2}$$

In addition, it is assumed that the sum of W_u is fixed for each individual, representing a limited amount of source activation (which has been associated with working memory capacity; see Anderson, Reder, & Lebiere, 1996; Daily, Lovett, & Reder, 2001); see Equation (3).

$$W = \sum_{u=1}^{N} W_u \tag{3}$$

As before, we will focus on the effect at the auxiliary embedded verb of the example (1) of Nicenboim et al. (submitted). At the auxiliary embedded verb (*hatte*, "had"), a dependency between this verb and the subject needs to be created via the retrieval of a singular nominative-marked NP (e.g., "the philanthropist"). For this retrieval, there are two competitors, an accusative-marked NP (e.g., "the assistant(s)") and a genitive-marked NP (e.g., "the director(s)"). These NPs are singular in the high interference condition and plural in the low interference one. For simplicity we will focus on three retrieval cues: (i) NP, which is a shared feature on both experimental conditions, (ii) nominative, which is a feature that uniquely distinguishes the target, and (iii) singular, which is shared with the competitors in the high interference condition, but uniquely distinguishes the target in the low interference condition.

In this setting, the spreading activation of the target for the low interference conditions, $S_{T|LI}$, described in Equation (4), would be higher than for the competitors, $S_{C|LI}$, and the spreading activation of both competitors should be described by the same Equation (5). The spreading activation of the target should be higher because regardless of the weights of the cues, W_u, it contains two strictly positive terms that are absent for the

competitors: the weighted strength of association of the nominative cue, $W_{nominative} \cdot s_{nominative}$, and of the singular cue, $W_{singular} \cdot s_{singular|LI}$, while both strengths of association are not shared and amount to the maximum associated strength, MAS; see Figure 2.

$$S_{T|LI} = W_{NP} \cdot \underbrace{s_{NP}}_{<MAS} + W_{nominative} \cdot \underbrace{s_{nominative}}_{MAS} + W_{singular} \cdot \underbrace{s_{singular|LI}}_{MAS} \quad (4)$$

$$S_{C|LI} = W_{NP} \cdot \underbrace{s_{NP}}_{<MAS} + W_{nominative} \cdot 0 + W_{singular} \cdot 0 \quad (5)$$

where the letter before the pipe, |, indicates whether it is the target, T, or the competitor, C, and the letters after the pipe, |, indicate whether it is under high, HI, or low, LI, interference.

Furthermore, the total activation, as described in Equation (1), is also affected by the noise component, ϵ, the history of usage or how much a representation decayed, B_c, and in the case of the competitors, it is also affected by a penalty component, P_c that quantifies the cost of a mismatch with the retrieval cues. This means that on average the activation of the target will exceed the activation of the competitors. Since in ACT-R the chunk with the highest level of activation is retrieved, this will mean that the target will be retrieved more often than any of the competitors. Regardless of which chunk is retrieved, the level of activation of the retrieved chunk determines the latency of the retrieval by Equation (6). This equation also ensures that a higher level of activation corresponds to a faster retrieval, which is scaled by F. To account for the absence of unrealistically long latencies and for cases where no chunk is retrieved, when the activation falls below a certain threshold, the retrieval fails (Lebiere, Anderson, & Reder, 1994).

$$Latency = F \cdot e^{-max(A_1,...,A_N)} \quad (6)$$

For the high interference conditions, the spreading activation of the target, $S_{T|HI}$ as shown in Equation (7), is still higher than the one of the competitors, $S_{C|HI}$ as shown in Equation (8). The difference between the spreading activation of target and competitors, however, is reduced because the singular cue is associated to both chunks in memory with strength $s_{singular|HI}$. Thus the strength of association, $s_{singular|HI}$, will be lower than the maximum associated strength, since here the singular cue does not uniquely identify the target. As before, however, on average, the activation of the target will exceed the activation of the competitors; see Figure 2.

$$S_{T|HI} = W_{NP} \cdot \underbrace{s_{NP}}_{<MAS} + W_{nominative} \cdot \underbrace{s_{nominative}}_{MAS} + W_{singular} \cdot \underbrace{s_{singular|HI}}_{<MAS} \quad (7)$$

$$S_{C|HI} = W_{NP} \cdot \underbrace{s_{NP}}_{<MAS} + W_{nominative} \cdot 0 + W_{singular} \cdot \underbrace{s_{singular|HI}}_{<MAS} \quad (8)$$

Because the strength of association between singular cue and the target in the high interference condition, $s_{singular|HI}$ in Equation (7), is smaller than the one in the low interference condition, $s_{singular|LI} = MAS$ in Equation (4), it follows that $S_{T|HI} < S_{T|LI}$. This means that on average, (when noise ϵ is canceled out) the activation of the target in the high interference condition would be smaller than in the low interference one. The higher activation of the target in the low interference condition leads to less errors, because the distributions of the activation of the target and competitors are further apart, and also leads to faster retrievals due to Equation (6). As a consequence, the activation-based account will predict higher accuracy and shorter reading times for the low interference condition in comparison with the high interference one.

Another less studied consequence of similarity-based interference is that when competitors are incorrectly retrieved in conditions of high interference, the process should take *shorter* time than when they are retrieved in low interference conditions. This is because the spreading activation of the competitors for high interference, $S_{C|HI}$, has the

term $W_{singular} \cdot s_{singular|HI}$, which is absent from the low interference conditions; compare Equation (5) with (8).

To sum up, in similarity-based interference the following inequalities derived from Equations (4)–(8) should hold:

$$A_{T|HI} > A_{C|HI} \tag{9}$$

$$A_{T|LI} > A_{C|LI} \tag{10}$$

$$A_{T|HI} < A_{T|LI} \tag{11}$$

$$A_{C|HI} > A_{C|LI} \tag{12}$$

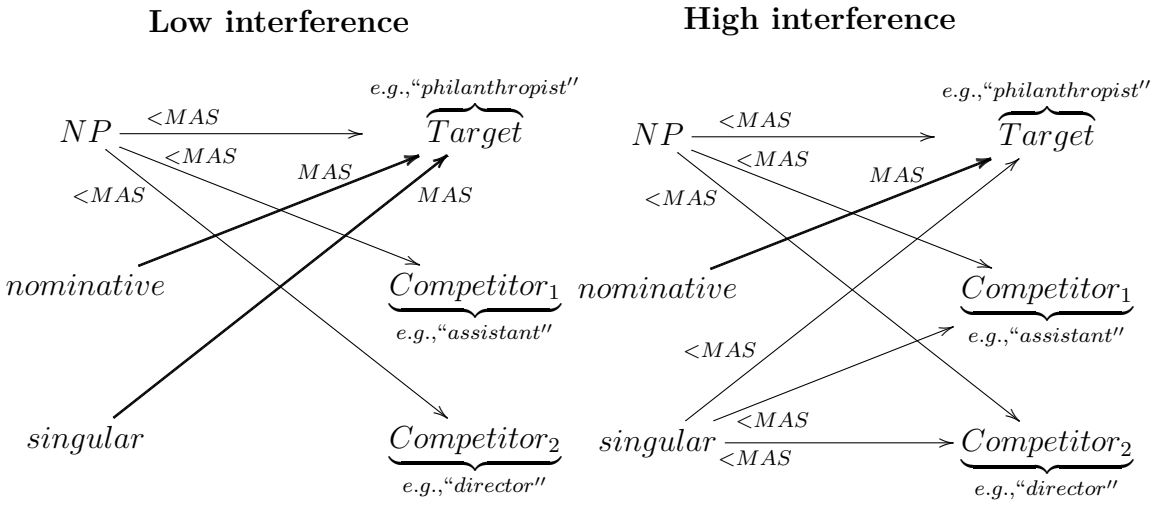

Figure 2. Graph showing the associations between the cues NP, nominative, and singular, and the target and two competitor NPs. The width of the arrow represents the strength of association, but this strength is weighted, so two identical associations may assign different amount of activation depending on the cue to which they belong.

The activation-based model as a lognormal race. In order to assess the fit of the activation-based account to the experimental data, we implemented it as a lognormal race of accumulators with a shift parameter and a single variance for the noise of all accumulators (Rouder et al., 2014), as explained further below. The retrieval in ACT-R can be thought of as a decision processes, where target and competitors stored in memory

accumulate evidence until the first chunk reaches a certain value and is retrieved. Activation can be linked to evidence by assuming that it represents the rate of its accumulation (in a similar way as assumed by van Maanen et al., 2011). This is so because activation in ACT-R represents the probability of the retrieval, and it is boosted by processes that increase evidence such as matching cues, previous retrievals and so forth, while it is penalized by processes that decrease the evidence such as mismatching features and decay.

Although in ACT-R the latency of the retrieval is only relevant for the chunk that was actually retrieved, as shown by Equation (6), our implementation assumes that there is a *potential* retrieval time, or *finishing time* in the race, t_c for each candidate c in memory. This is the time it would have taken for the chunk to be retrieved given its activation (scaled by the parameter F) :

$$t_c = F \cdot e^{-A_c} = e^{-A_c + log(F)} \tag{13}$$

The race model is implemented in the following way: Since the noise component in A_c is assumed to be normally distributed (Lebiere et al., 1994) and affects the activations of all the chunks to the same extent,[3] for each trial, the finishing time of each chunk, t_c, is sampled from a lognormal distribution with the same standard deviation σ, and the fastest chunk in a given trial would be the one retrieved (i.e., the chunk with the lowest t_c in a given trial).

$$t_c \sim e^{normal(-\mu_c + log(F), \sigma)} \Rightarrow \log(t_c) \sim normal(-\mu_c + log(F), \sigma) \tag{14}$$

$$\Leftrightarrow t_c \sim lognormal(-\mu_c + log(F), \sigma) \tag{15}$$

$$\tag{16}$$

[3]The noise component is sometimes approximated to have a logistic distribution for convenience, see, for example Lebiere (1999).

where

$$A_c = \mu_c + \epsilon \tag{17}$$

$$\epsilon \sim normal(0, \sigma) \tag{18}$$

The only observable data for every trial are (i) the answer of the comprehension question at the multiple choice task, w, (which we assume that when the question asks about the subject of the embedded verb, the answers correspond to the chunk retrieved from memory modulo offline distractions, i.e., the winner of the race), and (ii) the reading times at the site of the retrieval (the auxiliary verb "had.sg"). The reading times will include the retrieval time, $t_{c=w}$, of the "winner" chunk, and the time taken for other processes. Given the evidence that distributions of reaction times are shifted (Rouder, 2005; Nicenboim et al., 2016), we assume a lower bound, ψ, which represents changes in peripheral aspects of processing, such as encoding or motor execution (Rouder, 2005). We also account for other aspects of processing (e.g., lexical access) with a parameter p. For simplicity and to achieve a realistic fit to the data, p is added to the exponential factor (this ensures that the reading time variance increases with means, which is a general property of reaction times that seems to hold across many paradigms; see Rouder, Tuerlinckx, Speckman, Lu, & Gomez, 2008).

$$RT_{c=w} \sim \psi + lognormal(-\mu_c + log(F) + p, \sigma) \tag{19}$$

Since the observed reading times, as shown in Equation (19), are associated with the maximum activation (on a specific trial), the observable data constrain the unobserved finishing times, $t_{\forall c, c \neq w}$, of the other non-selected choices which must be necessarily slower:

$$t_{\forall c, c \neq w} > t_{c=w} \tag{20}$$

Since we are not interested in the specific value of μ_c, F, or p, but in learning from the model (i) whether the retrieval process resembles a race of accumulators, and (ii) the effect of number interference on the target and competitors, we fit the reading times, RT, as a function of α_c and an arbitrary constant, b, such that $b - \alpha_c = -\mu_c + log(F) + p$. By setting b large enough (to 10, for example), we ensure that α_c is strictly positive for ease of interpretation: a higher positive number corresponds with a higher rate of accumulation. Rouder et al. (2014) show that without further assumptions, thresholds and accumulation rates cannot be disentangled in the lognormal race model. To estimate the thresholds in a lognormal race model, Heathcote and Love (2012) assumed that both the rate of accumulation ν and the thresholds η were lognormally distributed, so that the finishing times, y, were distributed in the following way $y \sim lognormal(\mu = \mu_\eta - \mu_\nu, \sigma = \sqrt{\sigma_\eta^2 + \sigma_\nu^2})$. If the thresholds are fixed at some arbitrary point b, then $\sigma = \sigma_\nu$ (since $\sigma_\eta = 0$) and the rate of accumulation $\mu_\nu = b - \mu$, thus we can interpret α_c as the rate of accumulation associated with each chunk, and we can rewrite Equation (19) in the following way:

$$RT_{c=w} \sim \psi + lognormal(b - \alpha_{c=w}, \sigma) \tag{21}$$

On every trial, l, we can estimate $\alpha_{l,c=w}$ from the observed RT, and we can constrain the possible values of $\alpha_{l,\forall c,c\neq w}$ from the values of $t_{l,\forall c,c\neq w}$ that could not be possible on a given trial.

Given that shifts vary across participants but tend not to vary with experimental manipulation (Rouder, 2005), we assume a certain shift for every participant i, while $\alpha_{l,c}$, the activation together with nuisance parameters, will vary by participant i and by experimental item j.

$$RT_{l,i,j} \sim \psi_i + lognormal(b - \alpha_{l,i,j,c=w}, \sigma) \tag{22}$$

In standard ACT-R, if the activation is below a certain threshold, T, the retrieval

fails with a latency, $F \cdot e^{-T}$ (Lebiere et al., 1994). To avoid a deterministic latency, we assign an accumulator to the possibility of failure, which acts as a noisy timer, and its timeout depends on its parameter $\alpha_{c=failure}$. In this way, the retrieval threshold can be also thought as a chunk that competes for activation (Van Rijn & Anderson, 2003).

Figure 3 summarizes the parameters and the process for two chunks. The lognormal race model is sometimes called a *ballistic* or *deterministic* race model because there is no within-choice noise (as in, for example, the drift diffusion model): once a rate of accumulation is set for a given accumulator, it will determine its time to the threshold; this is represented in the lower part of Figure 3 by the straight lines. This, however, does not make the process deterministic, since the rate of accumulation changes from trial to trial.

It is important to highlight that even though we model the ACT-R retrieval process with a shifted lognormal race model, a good fit of the race model does not automatically imply that ACT-R predictions regarding number interference were borne out. For that to happen the inequalities (9)–(12) must hold.

In the implementation of the activation-based model, the hierarchical structure of this model is embedded in each parameter α_c associated with each chunk c (including one representing the failure) that is allowed to vary by condition (high/low interference) and includes by-participants and by-experimental-items intercepts and slopes (which are also allowed to be correlated). This means that the model can account, for example, for an NP of a certain experimental item being more semantically plausible as a retrieval candidate than other NPs, by simply adjusting the by-item intercept of the accumulators associated with each NP. See Appendix A for the details of the Bayesian model.

The direct access model

Next, we present an implementation of a direct access model. This is a Bayesian hierarchical implementation of the cue-based retrieval model proposed by Van Dyke and McElree (2006) based on McElree and colleagues' previous work (e.g., McElree, 2000;

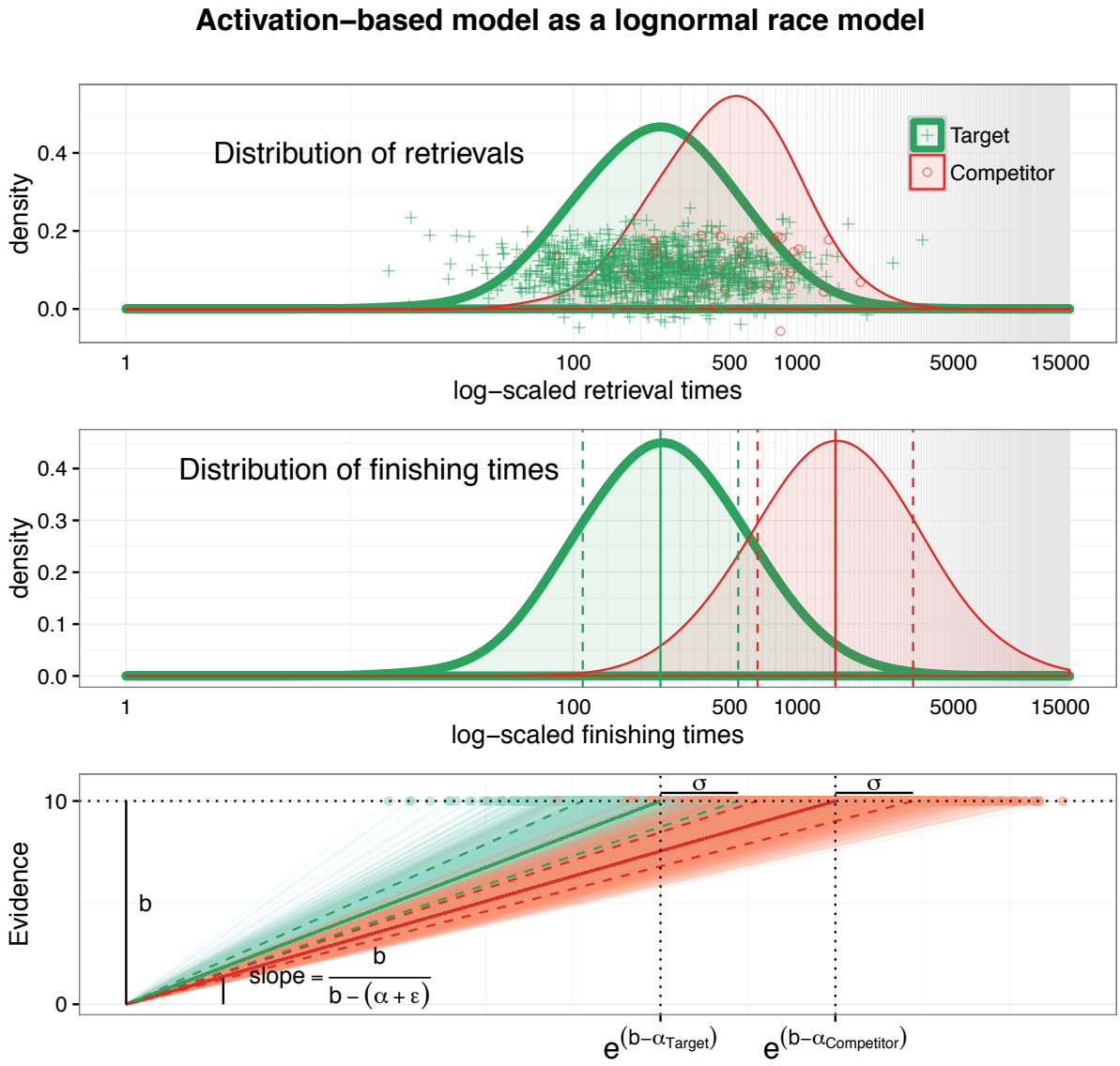

Figure 3. The figure depicts how the distribution of retrievals is generated from the activation-based model as a lognormal race model. The bottom figure depicts the parameters of the activation-based model, the full lines in green and red are the mean finishing times $t_{c=Target}$ and $t_{c=Competitor}$, and the broken lines are finishing times one standard deviation away from the mean. The middle figure shows the distributions of finishing times for target and candidate; since every chunk is associated a potential finishing time, t, both distributions have the same number of elements. The top-most figure shows the distribution of retrieval times (adding the shift parameter ψ would transform it to reading times); since only the winning chunks are retrieved, the distribution of retrieval times for targets has more elements than the one of the competitors. Notice that the retrieval times are faster than the finishing times: this is so because when a chunk has a very long finishing time in a given trial, it is very likely that its competitor will be faster.

McElree et al., 2003). Since the original model has not been implemented computationally, it is underspecified in some respects. We therefore made some assumptions in the model to spell these details out; these are described below.

For the direct access model, we assume as in Van Dyke and McElree (2006) that cues in the retrieval context are combined multiplicatively in a way similar to the proposed by the Search of Associative Memory (SAM) model (Raaijmakers & Shiffrin, 1980; Gillund & Shiffrin, 1984).[4] In Van Dyke and McElree (2006), the probability of retrieving the item or chunk (in ACT-R terms) c given N cues and O items in memory is defined by the following equation (where we follow a nomenclature as similar as possible to the one from ACT-R):

$$S_c = \prod_{u=1}^{N} w_u \cdot S_{c,u} \tag{23}$$

$$P(c) = \frac{S_c}{\sum_{k=1}^{O} S_k} \tag{24}$$

where S_c in Equation (23) is the total strength of chunk c which is defined as the product of the strength of association of the chunk c with each cue u weighted by w_u, and it is only limited to be positive.[5] One crucial difference between the multiplicative combination of cues and the additive one is that the former removes chunks that will never be retrieved from the "search set" of potentially retrievable chunks by assigning a strength of association close to zero. Thus Equation (24) is just the ratio of the total strength of item c divided by the sum of the total strength of the number of chunks.

Equation (24), however, does not allow for failed retrievals. In contrast with Van Dyke and McElree (2006), in the SAM framework (Raaijmakers & Shiffrin, 1980;

[4] Van Dyke and McElree (2011) are explicit, however, in that it could also be that cues are combined linearly, but with weights that are different enough so that certain cues, such as syntactic cues, have a more prominent role.

[5] This is sometimes called total activation of the image c (Raaijmakers & Shiffrin, 1980; Gillund & Shiffrin, 1984), but in order to distinguish it from activation from the previous account, we refer to it as "total strength."

Gillund & Shiffrin, 1984), it is assumed that Equation (24) represents the probability of sampling c from memory, but after it, there is a recovery process that is successful on a proportion of time that depends on the sum of the strength of association:

$$P_{recovery}(c) = 1 - e^{\sum_{u=1}^{N} w_u \cdot S_{c,u}} \tag{25}$$

In order to be able to account for failed retrievals, we will follow the distinction between sampling, recovery, and retrieval as it is present in the SAM framework; the entire process is shown in Figure 4 . We will focus on the probability of retrieval, P_r, of the target or one of the competitors in low or high interference, which entails sampling (with probability P) and recovery (with probability $P_{recovery}$). As before, for simplicity, we assume that there are three relevant cues for the retrieval of the subject, which explain interference effects in the example (1), namely, NP, nominative, and singular, and we derive the total strengths and probabilities of retrieval for low interference,

$$S_{T|LI} = (w_{NP} \cdot s_{NP}) \cdot (w_{nominative} \cdot s_{nominative}) \cdot (w_{singular} \cdot s_{singular}) \tag{26}$$

$$S_{C|LI} = (w_{NP} \cdot s_{NP}) \cdot (w_{nominative} \cdot s_{nominative*}) \cdot (w_{singular} \cdot s_{singular*}) \tag{27}$$

$$P_r(T|LI) = \frac{S_{T|LI}}{S_{T|LI} + 2 \cdot S_{C|LI}} \cdot P_{recovery}(T|LI) \tag{28}$$

$$P_r(C|LI) = \frac{S_{C|LI}}{S_{T|LI} + 2 \cdot S_{C|LI}} \cdot P_{recovery}(C|LI) \tag{29}$$

$$P_{failure} = 1 - (P_r(T|LI) + 2 \cdot P_r(C|LI)) \tag{30}$$

and for high interference.

$$S_{T|HI} = (w_{NP} \cdot s_{NP}) \cdot (w_{nominative} \cdot s_{nominative}) \cdot (w_{singular} \cdot s_{singular}) \tag{31}$$

$$S_{C|HI} = (w_{NP} \cdot s_{NP}) \cdot (w_{nominative} \cdot s_{nominative*}) \cdot (w_{singular} \cdot s_{singular}) \tag{32}$$

$$P_r(T|HI) = \frac{S_{T|HI}}{S_{T|HI} + 2 \cdot S_{C|HI}} \cdot P_{recovery}(T|HI) \tag{33}$$

$$P_r(C|HI) = \frac{S_{C|HI}}{S_{T|HI} + 2 \cdot S_{C|HI}} \cdot P_{recovery}(C|HI) \tag{34}$$

$$P_{failure} = 1 - (P_r(T|HI) + 2 \cdot P_r(C|HI)) \tag{35}$$

In order to allow for the retrieval of chunks that partially match the retrieval cues, we need to assume that there is a non-zero association between nominative and either accusative and genitive, $s_{nominative*}$, which is still smaller than the association between the cue nominative and the feature nominative, that is $0 < s_{nominative*} < s_{nominative}$. Similarly, there is a non-zero association, $s_{singular*}$, between the singular cue and the plural feature of the competitors in the low interference condition such that, $0 < s_{singular*} < s_{singular}$.

The direct access model as a mixture model. In order to account for differences in reading times, the direct access model assumes that, in some proportion of the cases, the parser is able to backtrack a misretrieval and to access the target candidate, taking some extra time. Thus the reading times associated with the correct retrievals are a mixture of directly accessed as well as backtracked retrievals as shown in Figure 5. This means that the probability of accessing a certain chunk, P' (which should be equivalent to the proportion of responses given at the multiple choice task modulo offline distractions), is affected by the probability of backtracking, P_b in the following way:

$$P'_r(T) = P_r(T) + (1 - P_r(T)) \cdot P_b \tag{36}$$

$$P'_r(C) = P_r(C) - P_r(C) \cdot P_b \tag{37}$$

$$P'_{failure} = P_{failure} - P_{failure} \cdot P_b \tag{38}$$

Regardless of the effect of P_b, which if large enough may produce a ceiling effect and hide the differences in accuracy between high and low interference conditions, we would expect the following relationships between the probabilities of retrieval (sample and recovery) to hold:

$$P_r(T|HI) > P_r(C|HI) \tag{39}$$

$$P_r(T|LI) > P_r(C|LI) \tag{40}$$

$$P_r(T|HI) < P_r(T|LI) \tag{41}$$

$$P_r(C|HI) > P_r(C|LI) \tag{42}$$

The core assumption of the direct access model is that retrieval takes the same time on average, t_{da}, regardless of the availability of the to-be-retrieved chunk. This is in contrast with the activation-based model, but also in contrast with the SAM framework of (Gillund & Shiffrin, 1984), where the retrieval time depends on the match between cues and features of the chunk. The implications for the direct access model is that there will be two different distributions of reading times: a distribution associated with the incorrect responses and a distribution associated with the correct ones. An incorrect response is given by a participant, only when the wrong chunk is retrieved and there is no backtracking and repair process; see Figure 5. In this case, the reading times at the retrieval site will only include the time needed for the direct access, t_{da}, together with the

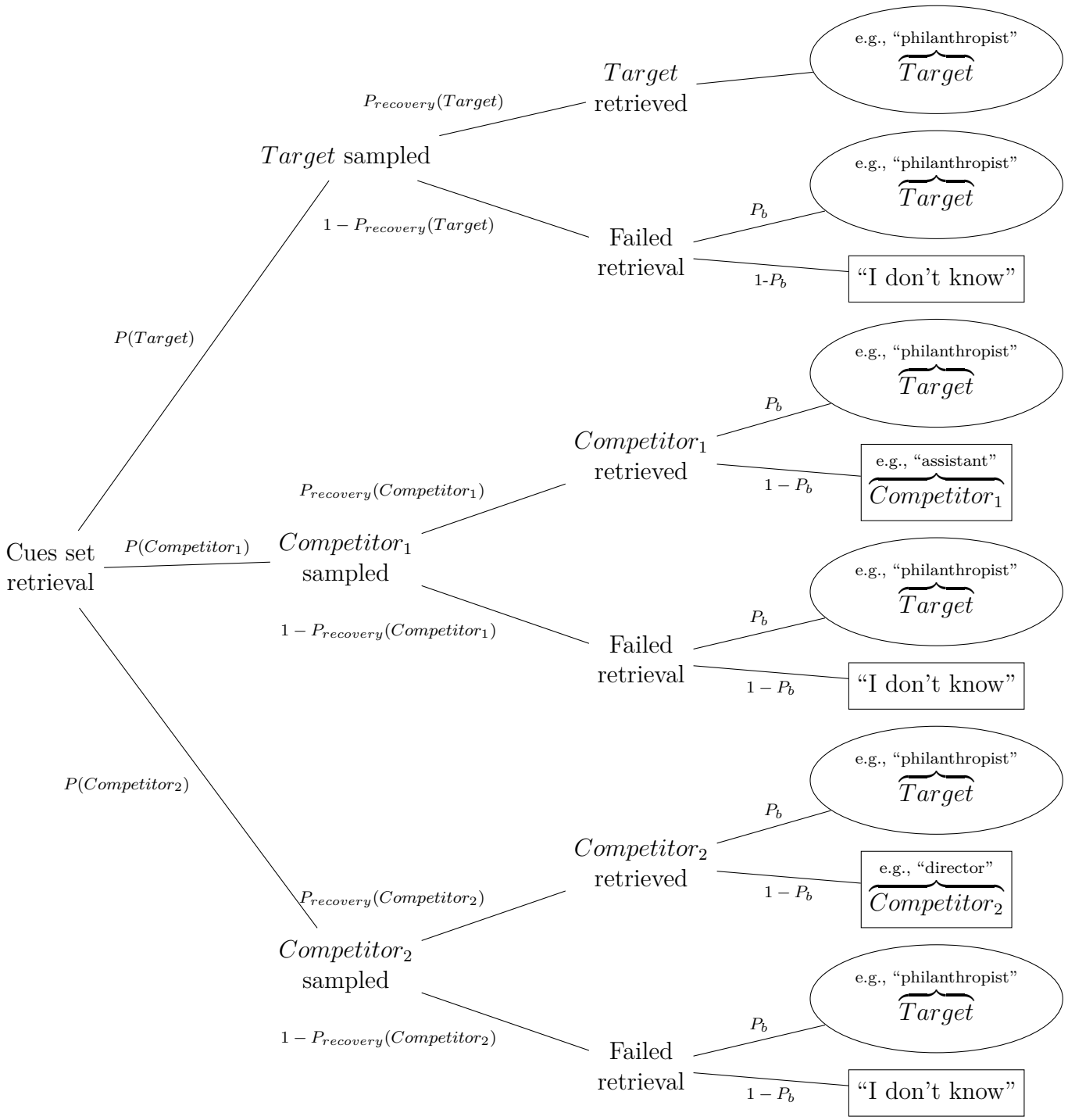

Figure 4. Graph showing how different responses are reached in the direct access model. Correct responses are inside ellipsis and incorrect ones inside rectangles.

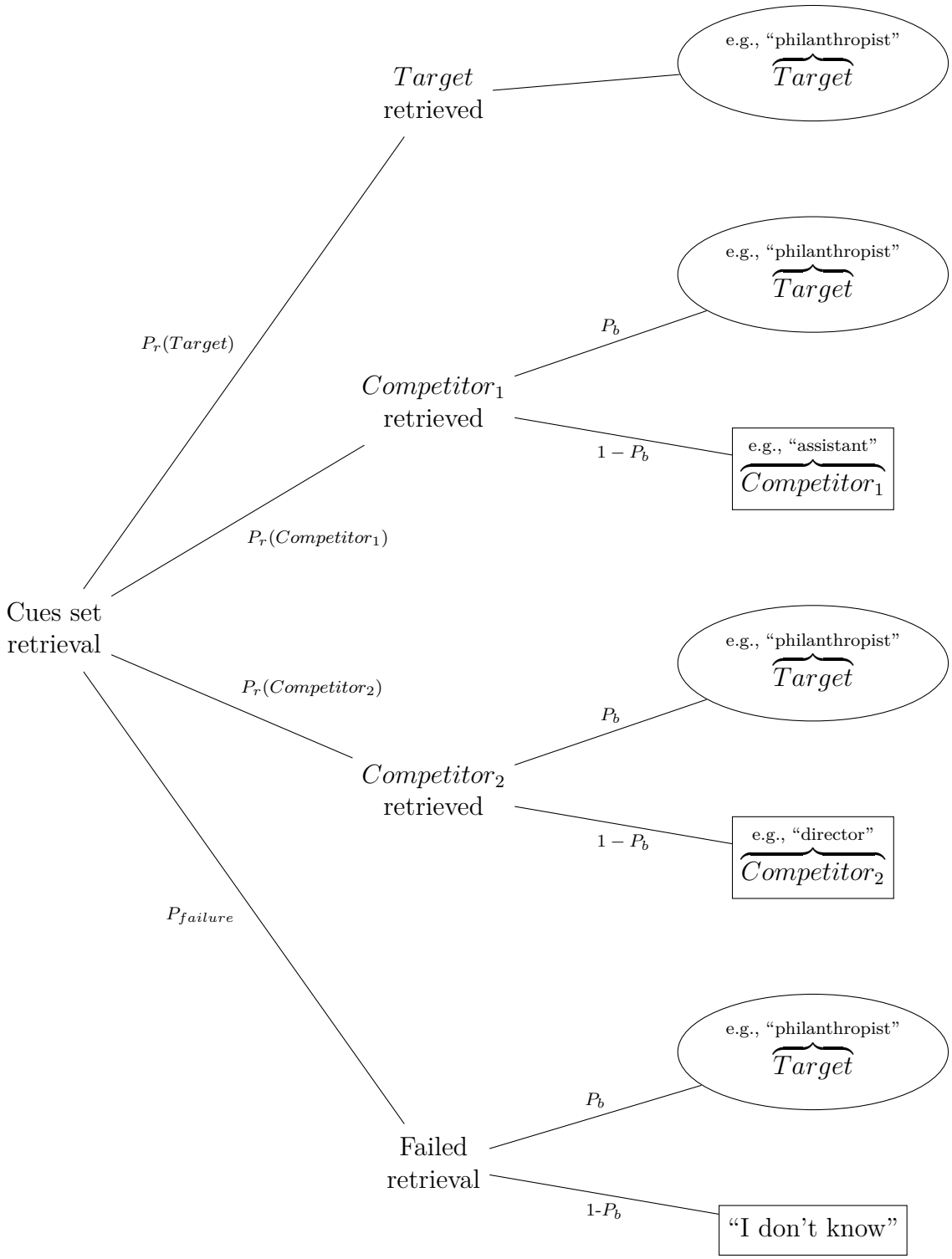

Figure 5. Simplified graph showing the different responses as a categorical distribution with correct responses inflated by the probability of backtracking, P_b. Correct responses are inside ellipsis and incorrect ones inside rectangles.

time taken for other processes, p, and normally distributed noise with standard deviation σ; crucially, this time is independent from the level of interference. We assume, as before, that the noise and the time taken for other processes are added to the location of the lognormal distribution; in addition, reading times are assumed to be shifted by some minimum amount ψ of time that represents the lower bound of the process. Thus we can assume that the reading times at the retrieval site for each trial l that are associated with an incorrect response (i.e, the retrieval of a competitor or a failed retrieval) have a shifted lognormal distribution, where the t_{da} depends on each participant i, and experimental item j, but not on the experimental condition or the identity of the chunk retrieved.

$$RT_{incorrect,l,i,j} \sim \psi_i + lognormal(t_{da,i,j} + p_{i,j}, \sigma) \tag{43}$$

For correct responses, the reading times depend on whether there is a repair process or not; see Figure 5. This entails that the distribution of reading times is a mixture of two components: The first one is associated with chunks correctly retrieved at the first attempt as shown in the first line of Equation (44), and it is identical to the distribution of incorrect responses. The second one is associated with incorrect responses that are backtracked and repaired, and it includes the direct access times, t_{da}, together with the time it takes to backtrack and do a reanalysis, t_b, as shown in the second line of Equation (44). (We derive the exact proportions in Appendix B.)

$$RT_{correct,l,i,j} \sim \psi_i + \begin{cases} lognormal(t_{da,i,j} + p_{i,j}, \sigma) & \text{, if the first try is correct} \\ lognormal(t_{da,i,j} + t_{b,i,j} + p_{i,j}, \sigma) & \text{, otherwise} \end{cases} \tag{44}$$

As with the activation-based model, we are not interested in p, and we define $T_{da,i,j} = t_{da,i,j} + p_{i,j}$. Support for this model would mean not only a good fit, but also that similarity-based interference would affect the probabilities of retrieval in the way shown by

the inequalities (39-42). We estimate the probability of each retrieval and the effect of interference on the retrieval probability using a multi-logit regression (or categorical distribution with the parameters on the logit scale). This is achieved by assigning a hierarchical structure to the parameters of the multi-logit regression which vary by condition (high/low interference) and include by-participants and by-experimental-items intercepts and slopes which have one correlation matrix for participants and one matrix for experimental items. Furthermore, we assign a hierarchical structure also to T_{da} and t_b, which are composed by by-participant and by-experimental-item varying intercepts. To allow for correlations between the direct access time and the backtracking time, we included also one correlation matrix for participants and one matrix for experimental items. This means that we assume that latencies should not be affected by retrieval probabilities. See Appendix B for more details.

Evaluation of the activation-based and direct access models

Application to data from a self-paced reading experiment

We fitted the models to a subset of the data from Nicenboim et al. (submitted). This work reports two self-paced reading studies investigating similarity-based interference with experimental items similar to (1). For the current study, we pooled the data of the 183 participants of both experiments, but keeping only the sentences with questions that queried the subject of the embedded verb, since cue-based retrieval models predict that interference will affect only retrievals where the cue is relevant. This left us with 20 sentences for each participant. For each sentence we used the time taken for reading the auxiliary verb *hatte* ("had.sg") and the response given at the multiple choice task; as we mentioned before we assume that the response given corresponds to the NP that was retrieved at the moment of parsing the auxiliary verb.

We fitted the models using *rstan* package (Stan Development Team, 2016a) in R (R Core Team, 2015) with four chains and 2000 iterations, half of which were the burn-in

or warm-up phase. In order to assess convergence, we verified that the $\hat{R}$s were close to one, and we also visually inspected the chains (Gelman, Carlin, et al., 2014). When needed, we also increased the maximum tree-depth and the adaptation parameter δ of the sampler to eliminate divergent transition and achieve convergence. We also verified that we could recover the parameters from the models using fake data simulation (Gelman & Hill, 2007).[6]

Posterior predictive checking

We use posterior predictive checking to examine the descriptive adequacy of the models (Shiffrin et al., 2008; Gelman, Carlin, et al., 2014, Chapter 6), that is, the observed data should look plausible under the posterior predictive distribution. The posterior predictive distribution is composed of 4000 datasets (one for each iteration) that the model generates based on the posterior distributions of its parameters. In other words, given the posterior of the parameters of the model (which are based on the current data), the posterior predictive distribution shows how other data may look like. Achieving descriptive adequacy means that the current data could have been predicted with the model. While passing a test of descriptive adequacy is not strong evidence in favor of a model, a major failure in descriptive adequacy can be interpreted as strong evidence against a model (Shiffrin et al., 2008). Thus, posterior predictive checking is an important sanity check to assess whether the model behavior is reasonable (see Gelman, Carlin, et al., 2014, for further discussion)

Given that the main difference between the activation-based model and the direct access model is in the way they account for the relationship between retrieval probability and latencies, for each of the 4000 datasets generated by the models, we calculate the means and .1-.9 quantiles of the reading times associated with each response, as well as the mean proportion of responses given. We represent this graphically using violin plots (Hintze & Nelson, 1998): the width of the violin plots represents the density of the

[6]Data and code can be downloaded from: www.ling.uni-potsdam.de/~nicenboim/code/code-data-retrieval-models.zip

predicted means (or quantiles). The observed mean (or quantile) of the data is represented with a cross. If the data could plausibly have been generated by the model, we would expect the crosses to be inside the violin plots.

Estimation of relevant parameters

In addition to fitting the data, the models include parameters that can be interpreted and can give support (or falsify) some assumptions of the effect of interference under the two presented models. We provide the estimates of some key parameters (or relations between key parameters) with their credible interval.

Cross-validation

We also compared the models using cross-validation, since the descriptive adequacy can also be achieved by a model that is too flexible and can generate too many different results. The idea behind cross-validation is to assess the accuracy the model would have in making predictions for new data, that is the expected predictive performance. The leave-one-out (LOO; Geisser & Eddy, 1979) method is a robust way to compare the expected predictive performance of the models (Vehtari & Ojanen, 2012; Gelman, Hwang, & Vehtari, 2014; Vehtari, Gelman, & Gabry, 2015). The basic idea of LOO is to split the data such that each training set used for estimating the parameters only excludes one observation, while one observation is validated. Then the estimate of the expected log pointwise predictive density ($\widehat{elpd}$) for a new dataset (i.e., the sum of the expected log pointwise predictive density of each observation) can be used as a measure of predictive accuracy for the N data points taken one at a time; $\widehat{elpd}$ can be transformed to deviance scale by multiplying it by minus two, providing a fully Bayesian alternative to AIC (Akaike Information Criterion; Akaike, 1974) or DIC (Deviance Information Criterion; Spiegelhalter, Best, Carlin, & Van Der Linde, 2002).

However, the robustness of LOO is achieved by fitting a model as many times as the number of observations, which is prohibitive in terms of time for complex models. In order

to reduce computation time, it is possible to use an approximation of LOO or to reduce the number of models to be fit. We first calculated $\hat{elpd}$ approximating LOO with Pareto smoothed importance sampling (PSIS-LOO; Vehtari & Gelman, 2015) with the R package *loo* (Vehtari, Gelman, & Gabry, 2016). However, PSIS-LOO can be affected by highly influential observations; the estimated shape parameter $\hat{k}$ of the generalized Pareto distribution can be used to assess the reliability of the estimates, with $\hat{k} > .7$ indicating an unreliable calculation of $\hat{elpd}$. Given that several $\hat{k}$ of the pointwise estimates were above 0.7 for the models presented here, we also provide $\hat{elpd}$ based on k-fold cross-validation (Vehtari & Ojanen, 2012), with k set to ten. We calculated the k-fold cross-validation by first splitting the data into 10 subsets (or folds) and then using each subset as the validation set, while the remaining data were used for parameter estimation. We partitioned the data into subsets by pseudo-randomly permuting the observations, and then systemically dividing them into 10 subgroups; we ensured that each group contained similar number of observations for each subject (this was meant to avoid the situation where most of the data of a certain subject is left out due to chance).

Results

Activation-based model.

Posterior predictive check. The posterior predictive check reveals that the model is inadequate for predicting some key characteristics of the data. Figure 6(a) shows that the model predicts shorter times for reading the auxiliary verb when the correct response is given and longer times for reading the auxiliary verb when an incorrect answer is given. In other words, the model underestimates the retrieval time of the correct dependent and overestimates the retrieval time of the competitor NPs, or the timeout. Figure 6(b) also shows a slight misfit for the predicted accuracy: the model tends to underestimate the proportion of correct responses and to slightly overestimate the proportion of incorrect ones. Furthermore, Figure 6(c) reveals that the fit is especially poor

for the second half of the quantiles.

Estimation of relevant parameters. The key parameters and relationships between parameters of the activation-based model are summarized in Figures 7(a) and (b). Figure 7(a) shows caterpillar plots of the posterior distributions for the rates of accumulation of evidence for each choice assuming an arbitrary threshold of 10. In the activation-based model, these parameters represent the mean activation (together with a common additive constant) of the target, competitor NPs, and in the case of the failure option, the activation represents the speed of the timeout. As assumed by the activation-based model, the activation of the target is higher than the activation of the competitors and of the failure. The figure shows that the activations of the chunks fit the inequalities (9-10), which indicate that the correct chunk should receive more activation on average than the competitor chunks.

As shown by Figure 7(b) the evidence for the effect of interference on the activation of the target and competitor NPs is rather weak. There is very weak evidence for interference decreasing the activation of the target ($\hat{\beta} = -0.02$, 95% CrI = $[-0.06, 0.03], P(\hat{\beta} > 0) = 0.23$) as predicted by Equation (11). Furthermore, even though interference effects should increase the activation of both competitors according to Equation (12), there is weak evidence that this might be the case for one of the competitors ($\hat{\beta} = 0.09$, 95% CrI = $[-0.04, 0.22], P(\hat{\beta} > 0) = 0.91$) and virtually no evidence for the second competitor ($\hat{\beta} = 0.03$, 95% CrI = $[-0.13, 0.19], P(\hat{\beta} > 0) = 0.66$).

Direct access model.

Posterior predictive check. The posterior predictive check reveals that, in contrast to the activation-based model, the direct access model is able to predict the main characteristics of the data fairly well. Figure 8(a) shows that the model is able to predict that reading times associated with correct responses are on average slower than the ones associated with incorrect ones, while Figure 8(b) shows that the model is able to predict fairly well the proportion of responses from the data. Furthermore, Figure 8(c) reveals that

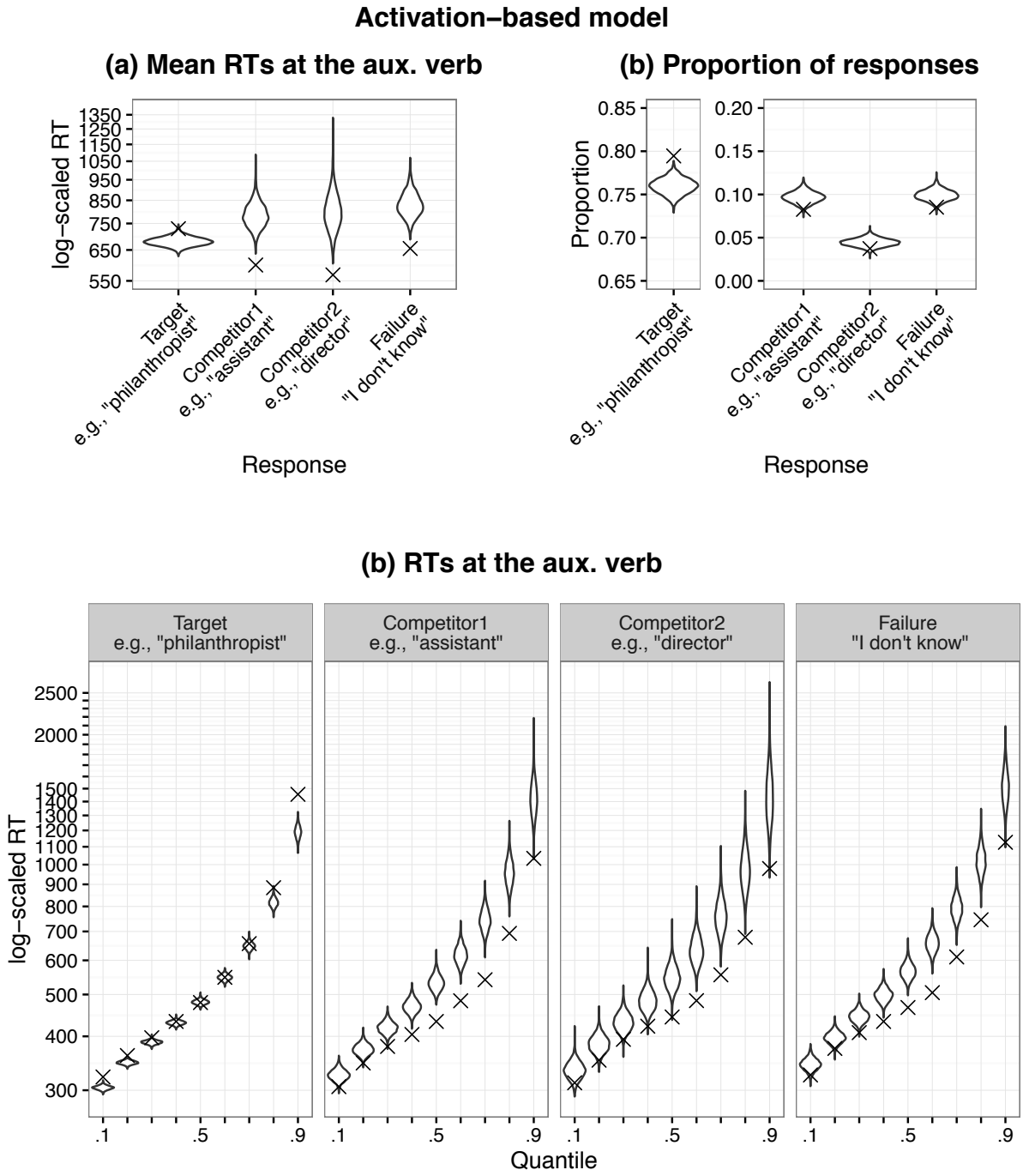

Figure 6. The top-most figure shows the fit of the mean reading times (RTs) for response (a) and proportion of responses (b) of the activation-based model. The width of the violin plot represents to the density of predicted mean RTs (a) and responses (b) generated by the model. The bottom figure (c) shows the fit of the .1-.9 quantiles of the reading times (RTs) for response of the activation-based model. The width of the violin plot represents to the density of predicted quantile generated by the model. The observed means and quantiles are represented with a cross.

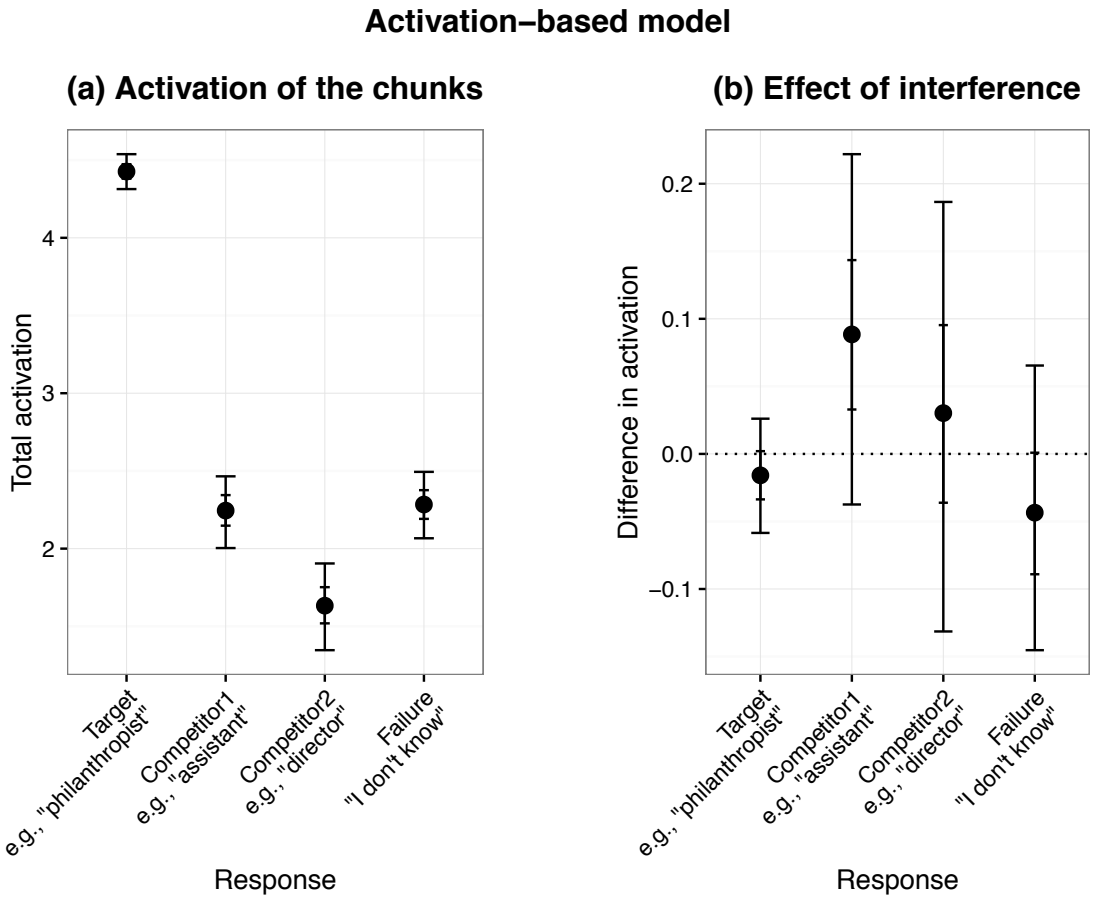

Figure 7. Mean activation of the different chunks assuming an arbitrary threshold of 10 (a), and mean difference between the activations of the chunks in high interference vs. low interference conditions (b). The outer error bars indicate 95% credible intervals while the inner error bars indicate 80% credible intervals.

the fit is generally good for the entire distribution of reading times.

Estimation of relevant parameters. The key parameters of the direct access model are: (i) the probability that each of the candidate NPs would be retrieved (as shown in Figures 9), (ii) the probability of backtracking (reported below), and (iii) the time needed for backtracking (reported below). Figure 9(a) shows caterpillar plots of the posterior distributions for the parameters that represent probability of retrieving each chunk from memory in order to build a dependency at the auxiliary verb. Figure 9(a) shows that the retrieval of the target is more likely than the retrieval of the competitors or

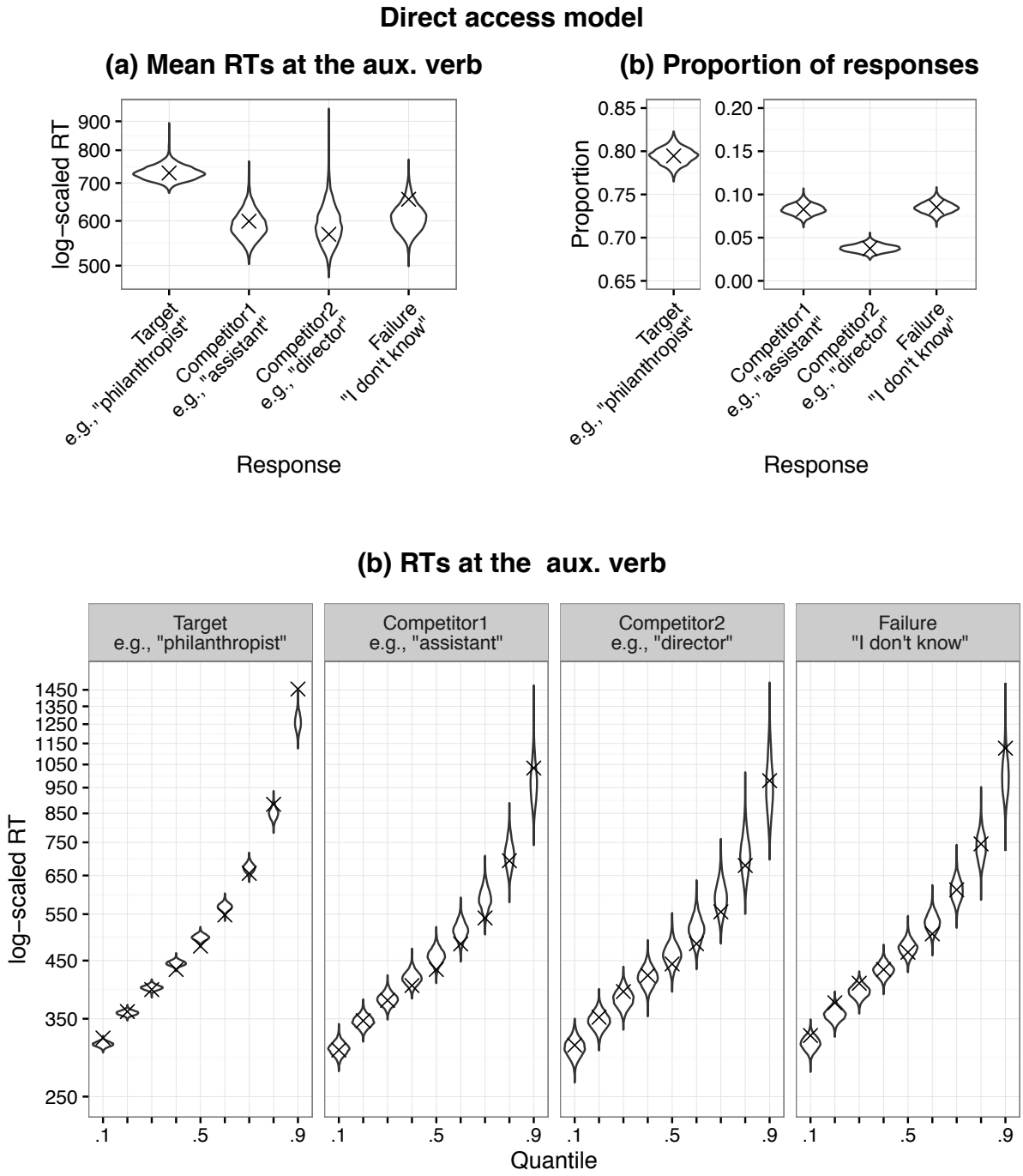

Figure 8. The top-most figure shows the fit of the mean reading times (RTs) for response (a) and proportion of responses (b) of the direct access model. The width of the violin plot represents to the density of predicted mean RTs (a) and responses (b) generated by the model. The bottom figure (c) shows the fit of the .1-.9% quantiles of the reading times (RTs) for response of the direct access model. The width of the violin plot represents to the density of predicted quantile generated by the model. The observed means and quantiles are represented with a cross.

the retrieval failure; this is in agreement with the inequalities (39-40). Notice that since the model assumes that backtracking is possible, after some trials the incorrect retrieval will be repaired. This means that the probability of retrieving a dependent is not the same as the proportion of times a responses was given in the multiple choice task. In fact, the model estimates that around half of the time that there is a misretrieval, it will be corrected ($\hat{\beta} = 0.48$, 95% CrI $= [0.4, 0.55]$). In addition, the model estimates that backtracking takes 119 ms, 95% CrI $= [25, 240]$ ms (after transforming it from log-scale).

As shown by Figure 9(b) the evidence for an effect of interference on the probability of retrieving the target or competitor NPs is not very strong. There is weak evidence for interference decreasing the probability of the retrieval of the target ($\hat{\beta} = -0.04$, 95% CrI $= [-0.11, 0.02], P(\hat{\beta} > 0) = 0.09$) and increasing the probability of incorrectly retrieving one of the competitors ($\hat{\beta} = 0.04$, 95% CrI $= [-0.01, 0.08], P(\hat{\beta} > 0) = 0.95$; and $\hat{\beta} = 0.02$, 95% CrI $= [-0.01, 0.04], P(\hat{\beta} > 0) = 0.88$); this is as predicted by Equations (41) and (42).

Cross-validation: activation-based vs. direct access models. In order to assess the compatibility of the models with the data, we compared how the models would generalize to an independent data set, that is, the pointwise out-of-sample prediction accuracy or $\widehat{elpd}$ of the models. We first used PSIS-LOO (Vehtari & Gelman, 2015) and then we verified the results with 10-fold cross-validation.

Comparing the models on PSIS-LOO reveals an estimated difference in $\widehat{elpd}$ of -119 ($SE = 28$) in favor of the direct access model in comparison with the activation-based model. However, given that there was a number of Pareto $\hat{k}$ larger than 0.7 (2.29% of the pointwise estimates for the activation-based model and 1.14% for the direct access model), we also compared the models with 10-fold cross-validation. This comparison also shows an advantage for the direct access model in comparison with the activation-based model, namely an estimated difference in $\widehat{elpd}$ of -110 ($SE = 28$) in favor of the direct access model.

Figure 10 shows for any given observation, whether one model has an advantage over

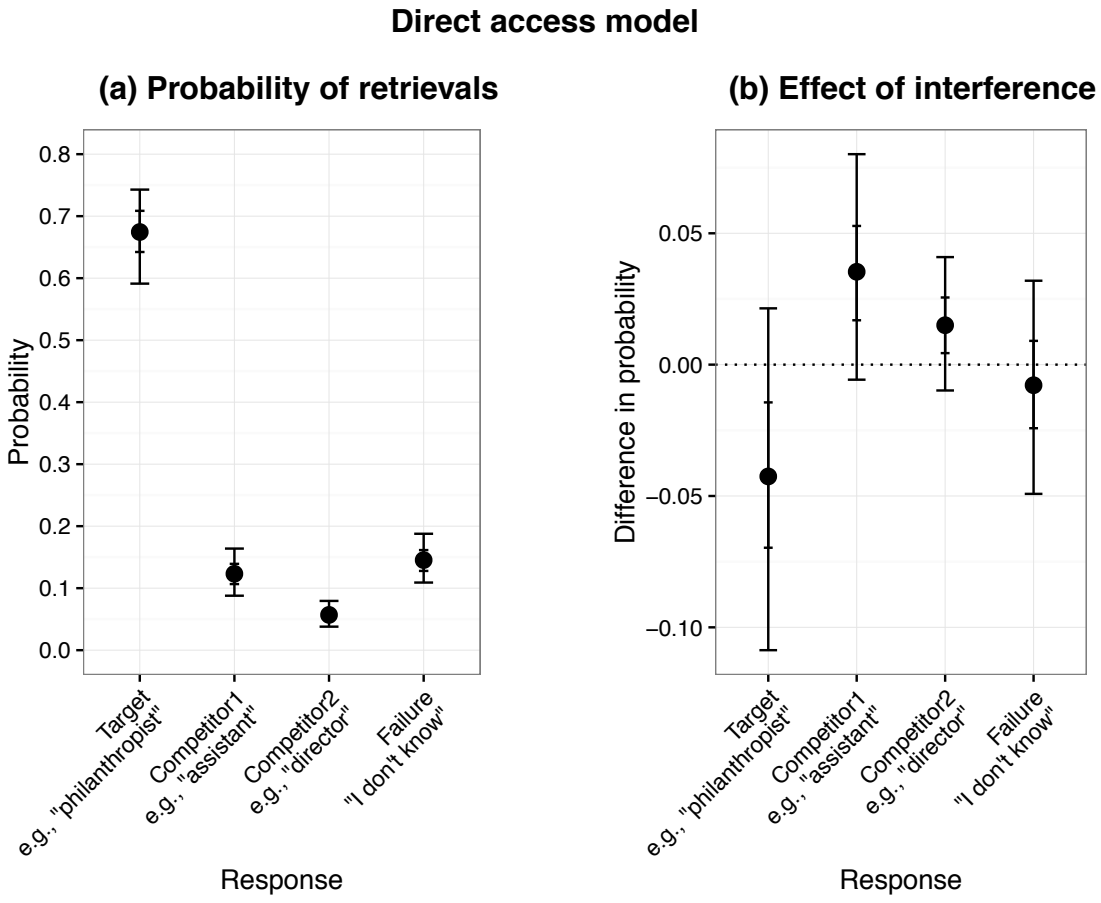

Figure 9. Mean probability of retrieval of the different chunks (a), and mean difference between the probabilities due to interference (b). The outer error bars indicate 95% credible intervals while the inner error bars indicate 80% credible intervals.

the other in its predictive accuracy. Since higher (or less negative) values of $\widehat{elpd}$ indicate a better fit for a model, observations that are further away from the dotted line correspond to data that are particularly better predicted by one model (and poorly by the other). This figure shows that the advantage of the direct access model is not due to some outlier observations, but mostly due to a high number of observations that fit slightly better under this model than under the activation-based one (this is the darker patch on the top right corner). Figure 11 shows the difference between the $\widehat{elpdf}$ of the two models for every observation corresponding to either a correct or an incorrect response. The figure shows that most of the advantage of the direct access model comes from reading times between

300 and 1000 ms (notice the darker patch above the zero dotted line). In addition, the direct access model has a clear advantage in predicting long reading times associated with correct responses and short reading times associated with incorrect ones, while the activation-based model has an advantage in predicting short reading times for correct responses and long reading times for incorrect ones.

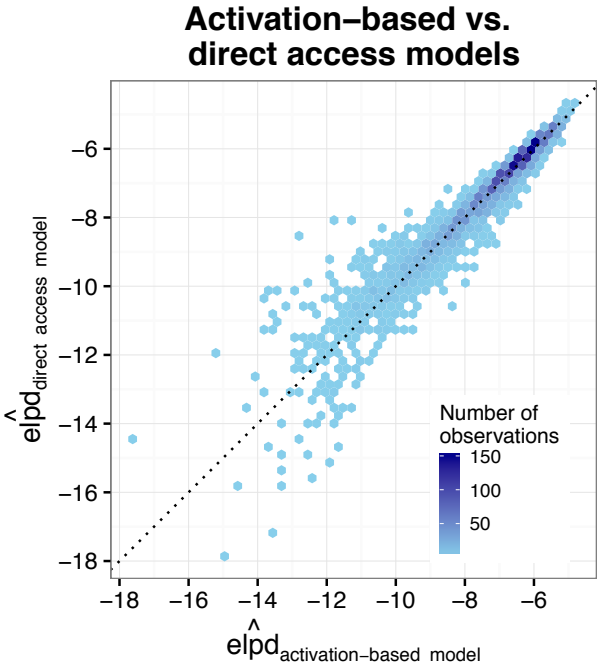

Figure 10. Comparison of the activation-based and direct access models in terms of their predictive accuracy for each observation. Each axis shows the expected pointwise contributions to 10-fold cross-validation for each model ($el\hat{p}df$ stands for the expected log pointwise predictive density of each observation). Higher (or less negative) values of $el\hat{p}df$ indicate a better fit. Darker cells represent a higher concentration of observations with a given fit.

Discussion

The evaluation of the activation-based and direct access models reveals two sets of findings: one relates to the effect of interference on the key parameters of the models, and other the relates to their validity as models of retrieval in sentence comprehension.

Regarding the effect of interference on the parameters of the models, the results show that interference affects the parameters as expected, but some of the posteriors include a

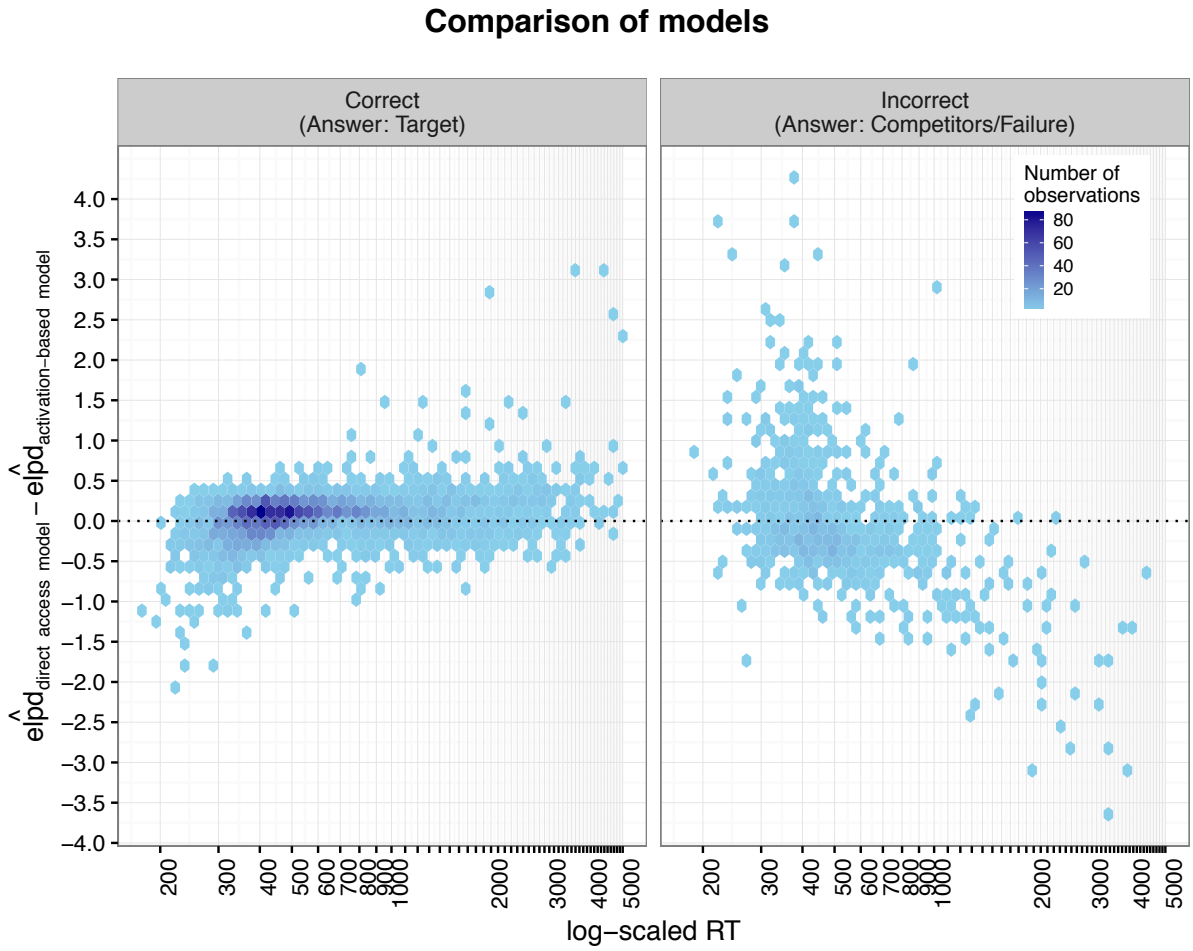

Figure 11. Comparison of the activation-based and direct access models in terms of their predictive accuracy for each observation depending on its log-transformed reading time (x-axis) and accuracy (left panel showing correct responses, and the right panel showing any of the possible incorrect responses). The y-axis shows the difference between the expected pointwise contributions to 10-fold cross-validation for each model ($\hat{elpdf}$ stands for the expected log pointwise predictive density of each observation); that is, positive values represent an advantage for the direct-access model while negative values represent an advantage for the activation-based model. Darker cells represent a higher concentration of observations with a given fit.

large degree of uncertainty. Given the small effect size on reading times in the experimental study of Nicenboim et al. (submitted), and given that we used a subset of the original data, this is not surprising. However, this serves as a sanity check that confirms that both models can work in principle and that experimental findings can produce the expected effects on the parameters of the models.

For the activation-based model, the underlying activation of the target NP was, as expected, clearly larger on average than the one of the competitors and the one associated with the timeout. The parameters that correspond to the effect of interference on activation, however, provided very weak evidence that interference decreases the activation of the target and increases the activation of the competitors.

Similarly for the direct access model, the underlying probability of retrieving the target was clearly larger than the one of the competitors and the one associated with the failure of the retrieval process. The model estimated that approximately half of the time ($\hat{\beta} = 0.48$, 95% CrI $= [0.4, 0.55]$) that a retrieval was incorrect, it was repaired to the correct retrieval in 119 ms, 95% CrI $= [25, 240]$ ms. This finding shows, as McElree (2000) suggested, that it is possible to account for differences in reading times that arise only from differences in probabilities of retrieval, if there is a repair process that takes more than a negligible amount of time. However, as with the activation-based model, the posterior distributions present only weak evidence that interference decreases the probability of retrieving the target and increases the probability of retrieving one of the competitors.

In order to evaluate the validity of the models for retrieval in sentence comprehension, we examined whether the models were able to fit the patterns found in the data using posterior predictive checks, and we compared their predictive accuracy using cross-validation. The posterior predictive checks of the activation-based and direct access models show clearly that some aspects of the data fit better under the direct access model than under the activation-based model. While we found that the reading times at the auxiliary verb associated with correct responses in the multiple choice task were on average slower than the reading times associated with incorrect responses, this pattern could only be captured by the direct access model. This is so because in the case of the direct access model, reading times associated with correct responses are assumed to be a mixture of fast direct-accessed retrievals and slower backtracked responses, while incorrect responses are assumed to be just direct-accessed wrong or failed retrievals. By contrast, in the case of the

activation-based model, reading times associated with correct responses are assumed to be faster on average than reading times associated with incorrect responses. The activation-based model assumes a race between the accumulation of evidence for the candidates to the retrieval, where the fastest item is the one retrieved. The particular characteristics of this race-between-accumulators model, which are motivated by ACT-R, include the assumption of a ballistic race (lack of fluctuations occurring during the accumulation process or within-choice noise; Brown & Heathcote, 2005) and the same variance parameter for all the accumulators (i.e., a single between-choice noise). Under this type of race, the correct responses, which are answered more frequently than the incorrect ones, will also be the fastest on average.

Even though reading times for correct responses were on average slower than the ones for incorrect ones, this was not the case for every observation. Model comparison using cross-validation shows that the advantage of the direct access model is based mainly on giving a better fit for reading times between 300 and 1000 ms, while the model is worst suited to predict fast reading times corresponding to correct responses and slow reading times corresponding to incorrect responses, which are better predicted by the activation-based model (see Figure 11).

The findings of cross-validation support the direct access model, but do not rule out others models that assume a race between accumulators of evidence for each retrieval candidate: the concept of activation determining retrieval latencies and accuracy may still be fruitful. It may be possible to explain the pattern in the data by including a mixture process in the race model, that is, if it is assumed, in a similar way as with the direct access model, that the reading times associated with the correct responses are a mixture of fast retrievals due to high activation together with repaired wrong or failed retrievals. However, a model like this would be too flexible for the data at hand and may present problems of identifiability (since it would be hard to estimate the activation of the non-retrieved candidates).

A closely related model that has been proposed to account for fast errors by Nicenboim et al. (2016) assumes that failed retrievals may take less time than completed retrieval. This is achieved by assuming that, when the activation is too low, the retrieval is aborted instead of waiting until the timeout is reached. However, this would only explain the fast failures ("I don't know answers" in the multiple choice task), but it would still leave fast retrievals of competitor NPs unexplained.

As we mentioned before, the activation-based model is based on a specific race model, namely the lognormal race model, which in turns is a very specific implementation of a model that assumes the sequential sampling of evidence for a decision (a class of models that includes the race of accumulators and random walk/diffusion models; for a review, see Ratcliff, Smith, Brown, & McKoon, 2016). There are other tasks that trigger error responses that are on average faster than correct responses and have been explained with sequential sampling models such as the drift diffusion model (Ratcliff & Rouder, 1998; Wagenmakers, Ratcliff, Gomez, & McKoon, 2008), or the linear ballistic accumulator (Brown & Heathcote, 2008). Fast errors can be captured by these models by assuming a lower threshold of evidence for the decisions. This will produce faster responses in general. However, if the initial bias to the responses is allowed to vary due to noise, the increase in speed will be larger for incorrect responses, because these responses mostly occur when there is a strong initial bias (Wagenmakers et al., 2008; Heathcote & Love, 2012). Even though the aforementioned models could account for the faster average reading times associated with incorrect responses, they would lose the close connection with the ACT-R framework that motivated our use of the lognormal race model.

Regarding the lognormal race model, its limitation is that if equal variance is associated with each accumulator, fast errors on average cannot be predicted because bias (distance) and rate of accumulation cannot be disentangled (Heathcote & Love, 2012; Rouder et al., 2014). Fast errors on average can be predicted, if the variance of the accumulators of the incorrect responses is larger than the one of the correct response.

Heathcote and Love (2012) propose that poorer matches may spread not only weaker activation on average but may also be noisier than stronger matches. This idea is also present in SAM framework of Gillund and Shiffrin (1984), which assumes slower reactions and more variability for poorer matches than for more precise matches. Figure 12 shows graphically how the distribution of correct and incorrect retrievals is generated from the activation-based model with different variances.

In the following section, we evaluate the activation-based model with different variances, one for the accumulator of target and one for the other accumulators, and we compare it with the direct access model.

Evaluation of the activation-based model with different variances

We evaluated the activation-based model with different variances using the same data as with the previous models. As before, we examined the descriptive adequacy of the model using posterior predictive checking, we estimate its relevant parameters, and finally we compared it with the direct access model using cross-validation.

The assumptions of the activation-based model with different variances are identical to the ones of the default activation-based model, except that the noise in the rate of accumulation of evidence of each chunk can have different variances. This means that the lognormal distributions associated with each activation have different scale parameters (which corresponds to the standard deviation of the associated normal distribution). Since all the competitors were retrieved only 21% of the time, for simplicity (and for improving the convergence of the models) we assumed only two variances, one for the lognormal distribution associated with the target, and one for the competitors or the failure timeout.

Results

Posterior predictive check. The posterior predictive check reveals that the activation-based model with different variances can capture the main characteristics of the data. Figure 13(a) shows that the model predicts a wide range of reading times associated

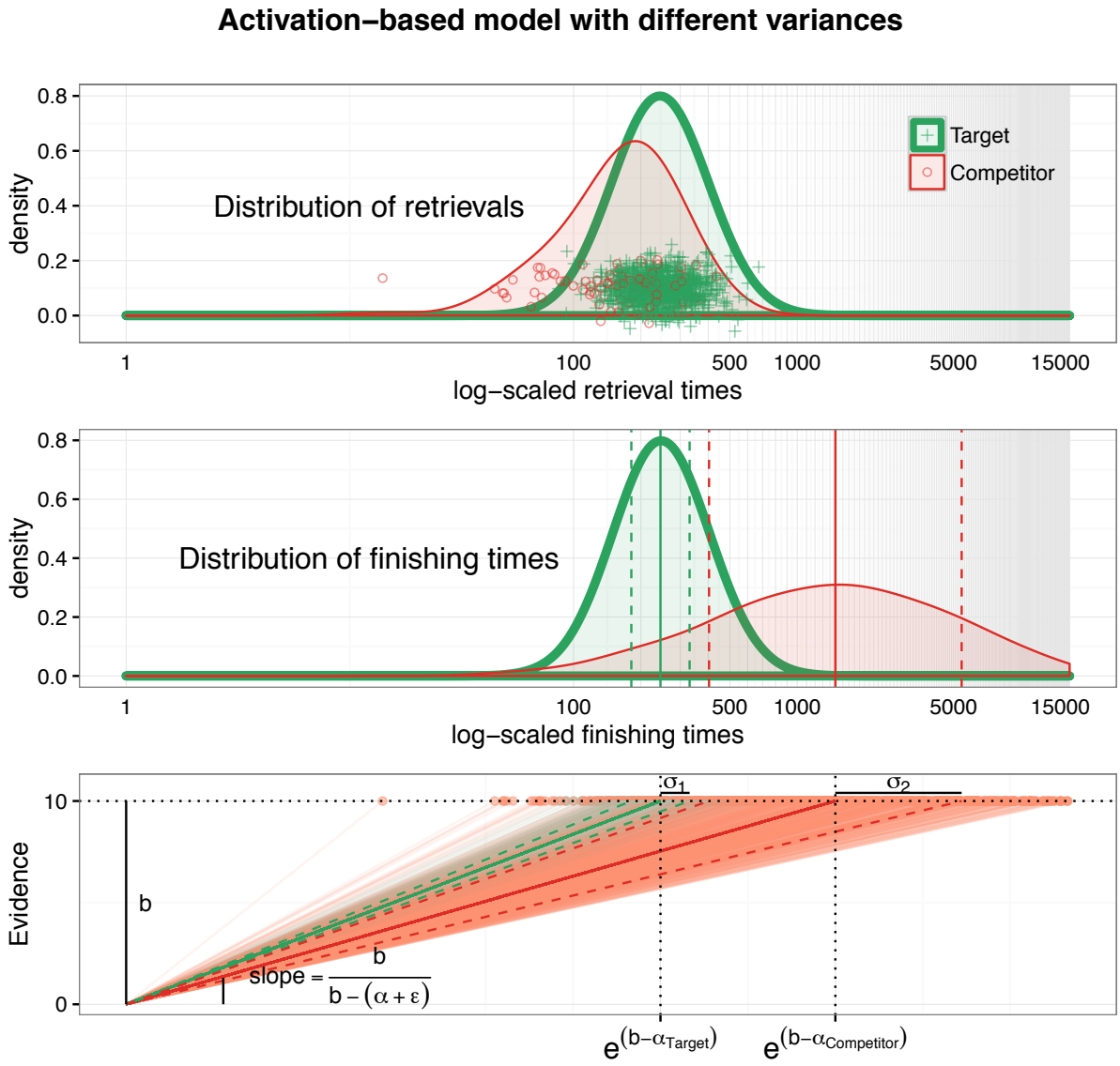

Figure 12. The figure depicts how the distribution of retrievals is generated from the activation-based model with different variances. The bottom figure depicts the parameters of the activation-based model, the full lines in green and red are the mean finishing times $t_{c=Target}$ and $t_{c=Competitor}$, and the broken lines are finishing times one standard deviation away from the mean. The middle figure shows the distributions of finishing times for target and candidate; since every chunk is associated a potential finishing time, t, both distributions have the same number of elements. The top-most figure shows the distribution of retrieval times (adding the shift parameter ψ would transform it to reading times); since only the winning chunks are retrieved, the distribution of retrieval times for targets has more elements than the one of the competitors. Notice that even though the finishing times for the competitors are slower on average than the ones of the targets (middle plot), the situation is reversed for the retrieval times (top-most plot).

with the incorrect responses, and most of the predicted reading times associated with incorrect responses are only slightly faster than the correct ones. Figure 13(b) shows that the model is able to predict the proportion of responses from the data. Figure 13(c) reveals that the fit is better for the first half of the quantiles, while for the second half of the quantiles the predicted data contains the observed quantiles, mainly because of the wide distribution of predicted reading times.

Estimation of relevant parameters. The estimation of the key parameters and the relationship between parameters of the activation-based model with different variances shows similar results to the ones in the default activation-based model. As in the default model, Figure 14(a) shows that the activation of the target is higher than the activation of the competitors and of the failure. In addition and also similarly to the case of the default model, Figure 14(b) shows that the evidence for the effect of interference in the target and competitor NPs is rather weak. There is very weak evidence for interference decreasing the activation of the target ($\hat{\beta} = -0.01$, 95% CrI $= [-0.05, 0.03], P(\hat{\beta} > 0) = 0.23$) and increasing the activation of the competitors (for the first competitor: $\hat{\beta} = 0.09$, 95% CrI $= [-0.05, 0.23], P(\hat{\beta} > 0) = 0.9$; and for the second competitor: $\hat{\beta} = 0.06$, 95% CrI $= [-0.11, 0.24], P(\hat{\beta} > 0) = 0.77$). As we mentioned before, the variance was allowed to be different for the correct and incorrect retrievals; Figure 15 shows that, as hypothesized, this allows the scale associated with the distribution of activations of the incorrect retrievals to be larger than the one associated with correct retrievals.

Cross-validation: activation-based model with different variances vs. direct access model. A comparison of the activation-based model with different variances and direct access model using 10-fold cross-validation shows that the estimates of $\widehat{elpd}$ are very similar, with a very small advantage for the activation-based model with different variances in comparison with the direct access model, namely an estimated difference in $\widehat{elpd}$ of -20 ($SE = 17$); while the advantage of the direct access model in comparison with the default activation-based model was of -110 with $SE = 28$.

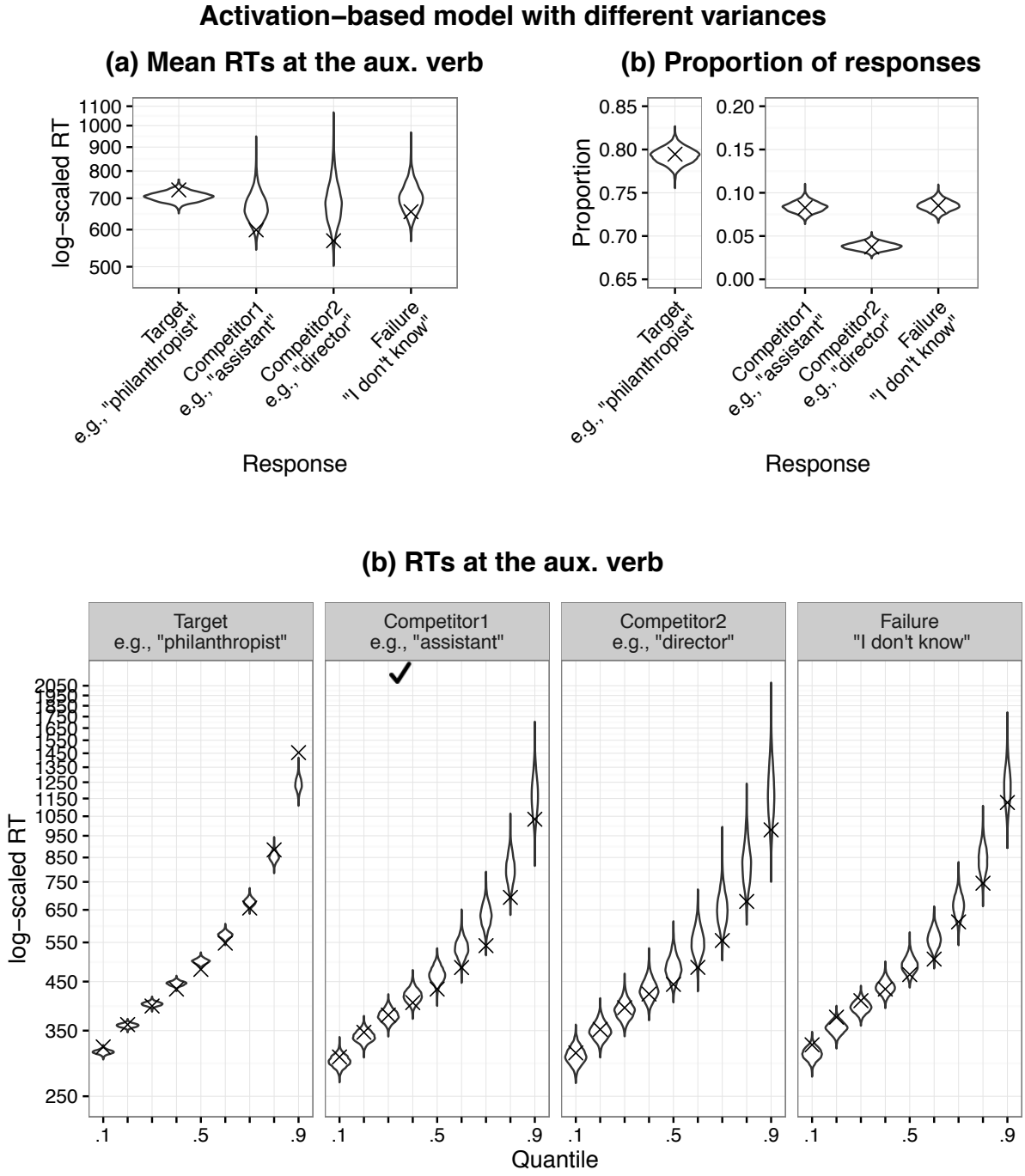

Figure 13. The top-most figure shows the fit of the mean reading times (RTs) for response (a) and proportion of responses (b) of the activation-based model with different variances. The width of the violin plot represents to the density of predicted mean RTs (a) and responses (b) generated by the model. The bottom figure (c) shows the fit of the .1-.9 quantiles of the reading times (RTs) for response of the activation-based model with different variances. The width of the violin plot represents to the density of predicted quantile generated by the model. The observed means and quantiles are represented with a cross.

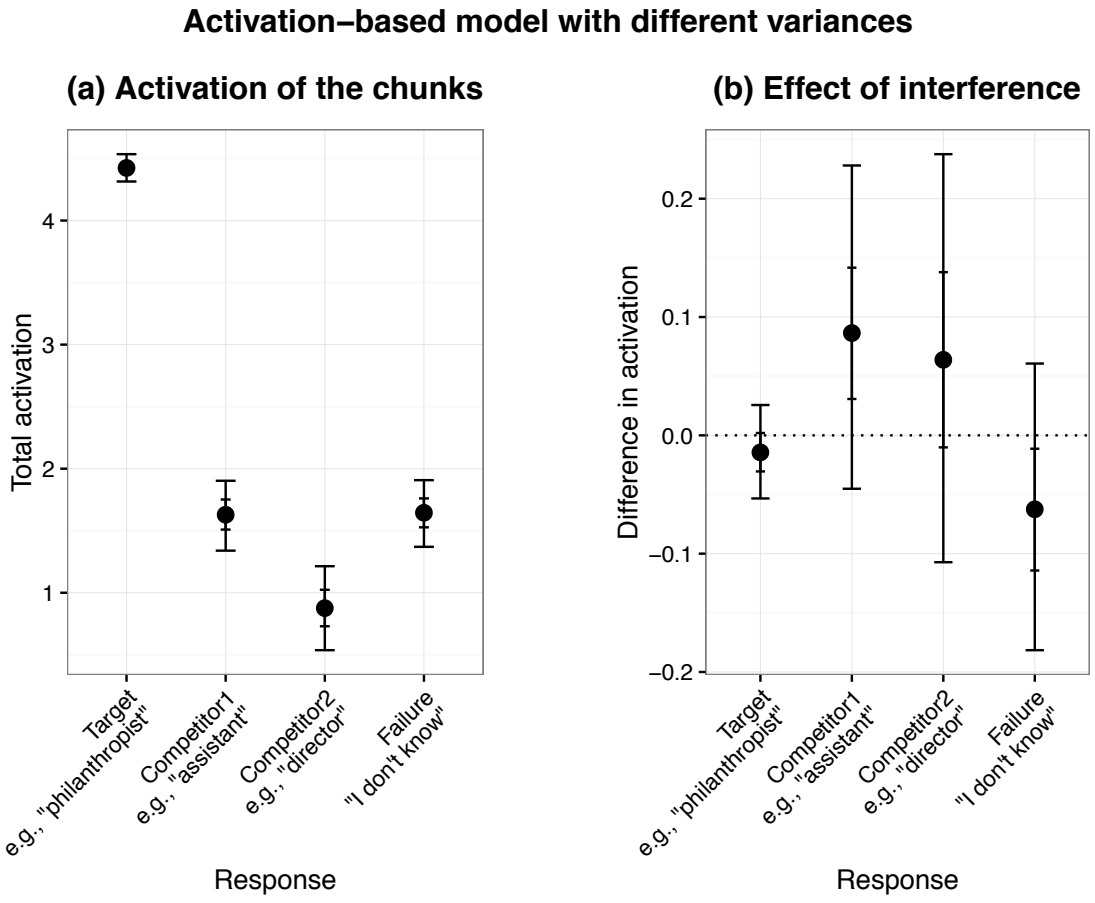

Figure 14. Mean activation of the different chunks assuming an arbitrary threshold of 10 (a), and mean difference between the activations due to interference (b). The outer error bars indicate 95% credible intervals while the inner error bars indicate 80% credible intervals.

Figure 16 shows that the predictive accuracy of the models is fairly similar with most of the observations being fit well by both of them. There are, however, some observations scattered at the bottom left corner of Figure 16, which favors the activation-based model with different variances. Figure 17 shows in blue cells the difference between the $elpd\hat{f}$ of both models for every observation corresponding to either a correct or an incorrect response; and in white cells the previous comparison (from Figure 11) of the default activation-based model and direct access model. Figure 17 shows that the difference between the activation-based model with different variances and direct-access model is

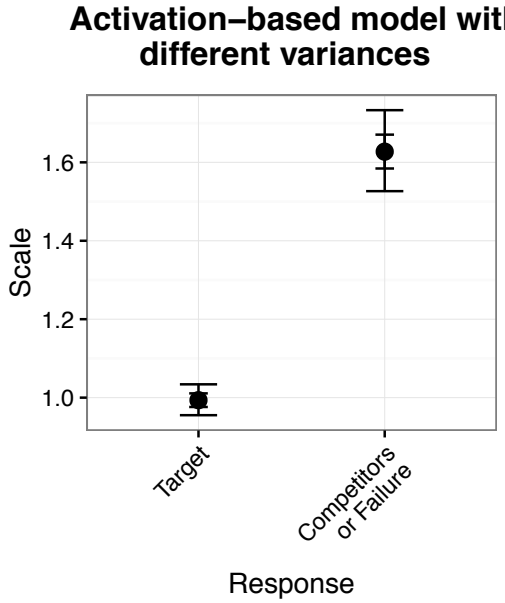

Figure 15. The figure depicts that the scale of the distributions of activations of the target chunk and of the competitors or timeout. The outer error bars indicate 95% credible intervals while the inner error bars indicate 80% credible intervals.

smaller than the difference between the direct access model and the (default) single-variance activation-based model. The main difference between the fits is that the activation-based model with different variances is able to account better for some fast and slow reading times associated with incorrect responses.

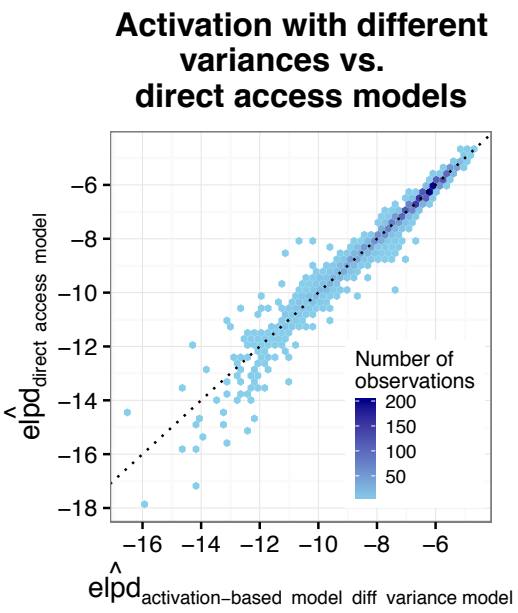

Figure 16. Comparison of the activation-based model with different variances and the direct access model in terms of their predictive accuracy for each observation. Each axis shows the expected pointwise contributions to 10-fold cross validation for each model ($el\hat{p}df$ stands for the expected log pointwise predictive density of each observation). Higher (or less negative) values of $el\hat{p}df$ indicate a better fit. Darker cells represent a higher concentration of observations with a given fit.

Discussion

The estimation of the relevant parameters of the activation-based model with different variances shows that the scale parameter associated with the distribution of activations for incorrect retrievals is larger than the one associated with correct ones, as it is necessary to account for fast errors. However, this did not change the predicted interference effect compared to the default activation-based model. Similarly, as with the default model, the parameters that correspond to the effect of interference on activation provide very weak evidence that interference decreases the activation of the target and increases the activation of the competitors.

Regarding the descriptive adequacy of the model, even though the inclusion of different variances improves the fit, the posterior predictive checks show more variation on

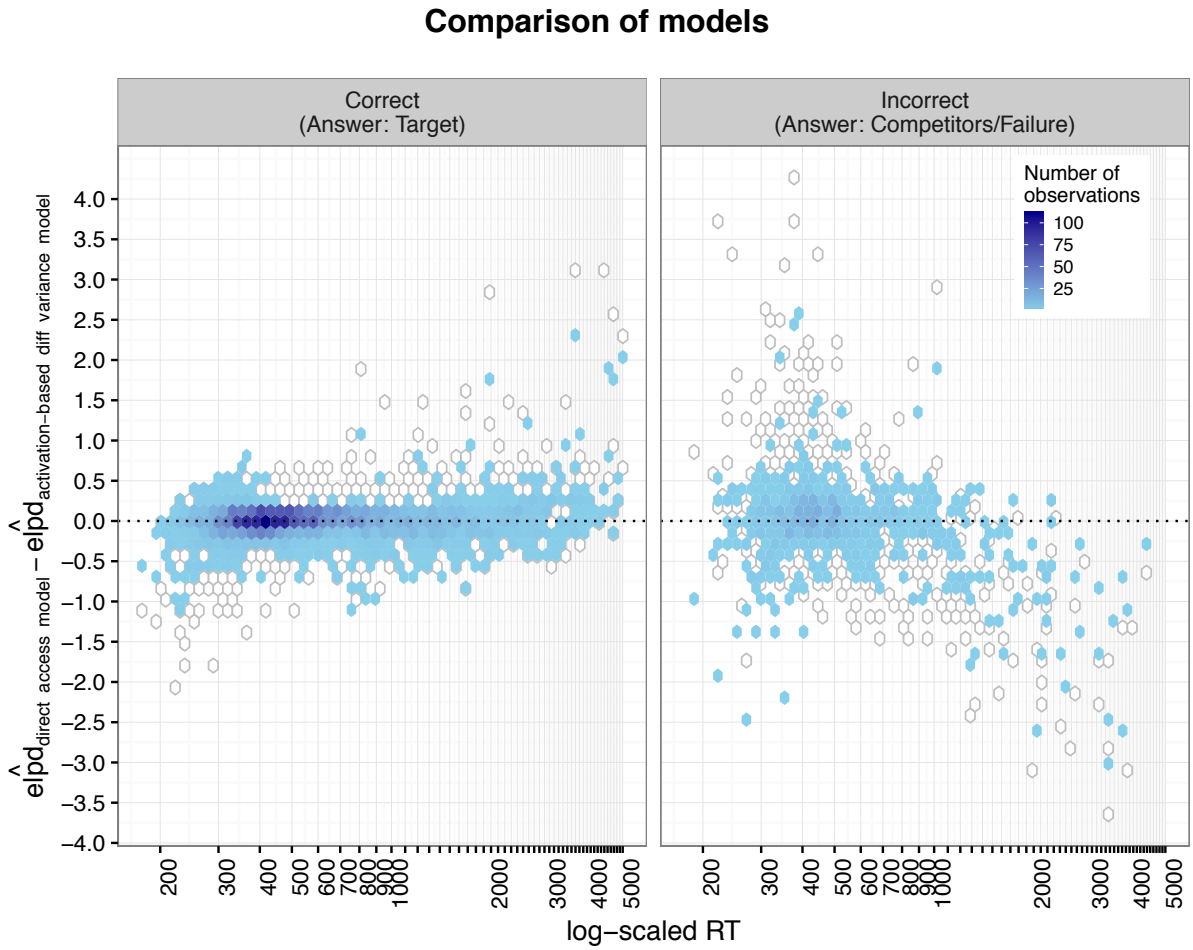

Figure 17. Comparison of the activation-based model with different variances and direct access model in terms of their predictive accuracy for each observation depending on its log-transformed reading time (x-axis) or accuracy (left panel showing correct responses, and the second panel showing any of the possible incorrect responses). The y-axis shows the difference between the expected pointwise contributions to 10-fold cross-validation for each model ($\widehat{elpdf}$ stands for the expected log pointwise predictive density of each observation); that is, positive values represent an advantage for the direct-access model while negative values represent an advantage for the activation model with different variances. Darker cells represent a higher concentration of observations with a given fit. The white cells show the comparison (shown earlier in Figure 11) of the default activation-based model with the direct access model.

the predicted reading times associated with incorrect responses for this model than for the direct access model. This is not necessarily a disadvantage, and it may indicate that the direct access model is more flexible and may be slightly overfitting the data, since these predictions are generated with the best estimates (and posterior distributions) to account

for the data. In fact, despite an apparent better fit for the direct access model, the estimates of predictive accuracy ($\widehat{elpd}$) are very similar with a very slight advantage for the activation-based model with different variances. In contrast with the difference between the fit of the original models (i.e., the default activation-based vs. the direct access model shown in Figure 11), the difference between the fit of the activation-based model with different variances and the direct access model is smaller (see Figure 17), with the new version of the activation-based model giving a better fit to some of the fast and slow reading times associated with incorrect responses.

This comparison shows that even though the inclusion of different variances for the accumulators does not imply a clear superiority over the direct access model, it is possible to account for the data with a model which is based on a race of accumulation of evidence.

General discussion

We evaluated two models that have been successful in explaining similarity-based interference in sentence comprehension: Lewis and Vasishth's (2005) activation-based model following ACT-R assumptions (Anderson et al., 2004) and McElree's (2000) direct access model. We also evaluated a third model, a variation of the activation-based model.

In order to compare the models we implemented them in a Bayesian hierarchical framework and we fit them to the data of Nicenboim et al. (submitted). Even though the activation-based model was already implemented computationally (Lewis & Vasishth, 2005), our implementation enabled us to go beyond simulations as they are usually done for this model (e.g., Vasishth & Lewis, 2006; Nicenboim et al., 2016), and fit the observations of an experiment accounting for variation coming from participants and experimental items. For the direct access model, we provide a first computational implementation which allowed us to derive precise and unambiguous predictions, which are fully transparent in our instantiation of the model. We first summarize our findings, and we then discuss the motivation of this work, the implications of the findings, and future work.

Our evaluation can be summarized in three main results: First, the underlying parameters of both models behave as expected under interference effects. However, the parameters showed a large degree of uncertainty in their posterior distributions. While this may be due to the small magnitude of the interference effect in the original experiment (Nicenboim et al., submitted), the findings confirm that, as expected, both models can in principle explain interference effects.

Second, we evaluated the validity of both models in predicting the reading times and accuracy patterns during retrieval. The posterior predictive checks and the comparison using cross-validation show that some aspects of the data fit better under the direct access model than under the default activation-based model. The data showed on average slower reading times associated with correct responses than with incorrect ones, and this pattern could be explained only by the direct access model. This suggests that the default activation-based model may not be flexible enough to accommodate patterns in the data that go beyond means between conditions.

Third, we show that by introducing a modification to the default activation-based model, namely, by assuming that the accumulation of evidence for the retrieval of incorrect items is not only slower but noisier, the new model can provide a fit as good as the one of the direct access model.

The importance of a formal evaluation of Lewis and Vasishth's (2005) activation-based model and McElree's (2000) direct access model lies in disentangling their predictions: Since both models assume that dependencies of non-adjacent elements are created via a content-addressable cue-based retrieval mechanism, they have been used almost interchangeably to explain interference effects (e.g. Van Dyke & McElree, 2006). For experiments that draw inferences from differences in means, these two models yield identical predictions for the inhibitory effect of similarity-based interference: namely, longer reading times at the retrieval of a dependent and/or a reduction of comprehension accuracy when several items share a feature associated with a retrieval cue. However, these models

are based on different underlying assumptions. The activation-based model follows ACT-R assumptions (Anderson et al., 2004); in this framework, the activation of the items in memory determines the retrieval accuracy and latency, and the activation of the target of retrieval is, in turn, adversely affected by interference. Crucially, latency and accuracy are not deterministic because activation fluctuates due to noise in the system. We show that this process can be seen as a lognormal race between accumulators of evidence with a single variance for all the accumulators, where activation represents the rate of accumulation of evidence. In contrast, the direct-access model assumes a model of memory where only the probability of retrieval can be affected by interference, while items take the same time to be retrieved (if they are not in the focus of attention as it is the case for non-local dependencies). In this model, differences in latencies are a by-product of the possibility of backtracking and repairing incorrect retrievals. These different assumptions lead to a different behavior in the relationship between reading times and response accuracy on a trial-level basis which cannot be examined by only comparing mean reading times or accuracy between conditions. While acceptability judgment tasks with speed-accuracy trade-off (SAT) allow a finer grain look at the reaction times and have been used to argue in favor of the direct-access model (see, for example, Van Dyke & McElree, 2011), there has been until now no computational evaluation of the model in reading for comprehension.

While the activation-based model uses the declarative retrieval module of ACT-R, which has been shown to be an empirically successful model (e.g., Anderson et al., 1998; Anderson & Reder, 1999; Van Rijn & Anderson, 2003), our findings show that its default implementation cannot account for wrong retrievals that were generally faster than the correct ones in our data. The model cannot account for this pattern because items in memory that match the retrieval cues will have higher activation on average than competitors that match the retrieval cues only partially. The higher activation on average leads in turn to faster retrievals on average. In contrast, the direct access model can successfully accommodate faster incorrect retrievals. This is done by assuming that reading

times associated with correct responses are generated from a mixture distribution of fast directly accessed correct retrievals at the first attempt together with slower repaired retrievals. Reading times associated with incorrect responses, in contrast, belong to a faster distribution of retrieval latencies of items that are directly accessed. It should be noted that this repair mechanism that explains slow correct retrievals could in principle be added to the activation-based model, but it would lead to an unidentified model. The direct access model, however, is able to account for the data with a very simple architecture that can integrate this repair mechanism. The direct access model assumes a bipartite architecture for retrieval (e.g., McElree & Dosher, 1989; McElree, 2006): Items within focal attention are accessed quickly, but all other items outside attention are accessed more slowly and with the same retrieval speed.

While the simple architecture of the direct access model may be preferred on grounds of parsimony, the activation-based model is compatible with a sequential sampling framework (such as the drift diffusion model: Ratcliff, 1978; the leaky competitive accumulator: Usher & McClelland, 2001; linear deterministic models: Heathcote & Love, 2012, among others) and has some possibly desirable characteristics. In the sequential sampling framework, decisions (such as which is the right dependent that needs to be retrieved) are considered a process of noisy accumulation of evidence, which has been shown to be compatible with the behavior of populations of neurons (e.g., Zandbelt, Purcell, Palmeri, Logan, & Schall, 2014). In addition, sequential sampling has been also linked to theories of optimality (Ratcliff et al., 2016; Summerfield & Tsetsos, 2015), which compare how an ideal agent would perform (given the levels of uncertainty in the stimuli) with the actual behavior of participants.

The sequential sampling framework could still be useful to explain retrieval, if we would assume that the retrieval process behaves similarly to other more complex accumulator models such as the linear ballistic accumulator (Brown & Heathcote, 2008). Thus, the fast incorrect retrievals on average could be captured by assuming a lower

threshold of evidence for the decisions. The initial bias to the candidates (which varies due to noise) would reduce the latencies for incorrect retrievals, since these mostly occur when there is a strong initial bias (Wagenmakers et al., 2008). However, this model loses the close connection with the ACT-R framework that motivated the lognormal race which underlies the activation-based model. In addition, given that the linear ballistic accumulator model is more complex than the lognormal race model, it is not clear whether its fit would be comparable to the fit of the direct access model. A potential future direction of this work would be to evaluate different plausible accumulator models as models of retrieval.

In the present study, we relaxed one of the assumptions of ACT-R to capture the patterns of the data: Here we assumed that the activation of chunks that match the retrieval cues only partially is not only lower but also noisier. This is translated into assuming different variances for the different accumulators. Heathcote and Love (2012) show that when the accumulators associated with incorrect responses have a larger variance than the accumulator of correct responses, the model can account for fast errors on average. For simplicity, we assumed one variance for the accumulator of the target, and one for the competitors and failure accumulators. While our study shows that this is enough to account for the pattern in the data, nothing would prevent all accumulators from having different variances.

Both the activation-based model with different variances and the direct access model showed equally good fit to the data. In order to investigate these models relative fit, future work should replicate the classical interference results (e.g., Van Dyke & McElree, 2006; Van Dyke, 2007; Van Dyke & McElree, 2011), while including reading times and questions probing the comprehension of the relevant dependencies.

In addition, there are other phenomena that the models could explain. These are: (i) the facilitatory interference effects found in ungrammatical sentences (Wagers, Lau, & Phillips, 2009), (ii) the ambiguity advantage in relative clauses (Traxler, Pickering, & Clifton, 1998) and the effect of task demands (Swets, Desmet, Clifton, & Ferreira, 2008),

and (iii) good-enough processing (Ferreira et al., 2002). Since some of the predictions of the activation-based model with different variances are not very intuitive, we provide an *R* script called *race-plot* using the *Shiny* package (Chang, Cheng, Allaire, Xie, & McPherson, 2016) that can help visualizing the predictions.[7]

Facilitatory interference. Wagers et al. (2009) noticed that the so-called number attraction effect in ungrammatical sentences such as (2), that is, the speedup in *are* in (2b) vs. (2a), could be accounted by Lewis and Vasishth's (2005) activation-based model.

(2) a. * The key$_{+sing}$ to the cabinet$_{+sing}$ are in the box.

b. * The key$_{+sing}$ to the cabinets$_{+plur}$ are in the box.

In sentences like (2), a cue-based retrieval mechanism would assume that a retrieval is initiated at the verb (*are*) with at least two retrieval cues: grammatical subject and plural. In sentence (2a), *the key* matches one of the retrieval cues, because it is the grammatical subject, but mismatches the plural cue. In sentence (2b), both nouns partially match the retrieval cues: *the key* matches the grammatical subject cue, while *the cabinets* matches the plural cue. An interesting prediction of the activation-based model (confirmed by experimental findings; see Jäger et al., submitted, for a meta-analysis) is that reading times are *faster* at the verb in (2b) than in (2a). This is so because a situation with no unambiguous match (both nouns are partial matches) leads to statistical facilitation (Raab, 1962), that is, an overall speedup when we examine mean reading times (facilitatory interference). For facilitatory interference in ungrammatical sentences, the predictions of the default activation-based model and the activation-based model with different variances are the same. This situation can be simulated using *race-plot* script mentioned before, by assigning arbitrary (but plausible) activations to the candidates to retrieval in (2a) and (2b): In (2a), *the key* (partial match) can be assigned an activation of 4 and *the cabinet*

[7]The application can be accessed in the browser with the following commands in R:

```
install.packages(c("dplyr", "tidyr", "ggplot2","cowplot","shiny")) #if needed
library(shiny) #load shiny
runUrl("http://www.ling.uni-potsdam.de/~nicenboim/code/race-plot.zip")
```

(no match) an activation of 2.5 (and $\sigma = 1.5$); this would result in a mean reading time of ≈832 ms. Notice that since the process is not deterministic, different simulations will show different retrieval times; the relationship between the conditions, however, should hold. In (2b), *the key* (partial match) can be assigned an activation of 4 and *the cabinets* (partial match) an activation of 3.5 (since they will not necessarily reach exactly the same activation); this would result in a faster reading time on average, namely ≈692 ms.

In contrast to the activation-based model, the direct access model would not predict a difference in reading times at the verb between (2a) and (2b). This is the case since increased reading times depend only on backtracking, which would only occur to repair an initially incorrect retrieval. In ungrammatical sentences with partial match, it is unclear how the repair would work, and why there would be more backtracking in (2a) than in (2b).

The predictions of the activation-based model, however, have not been investigated taking into account both reading times and comprehension. Even if a speedup compatible with facilitatory interference has been reported in the literature, the activation-based model would be accounting for facilitatory interference only if participants reach a different interpretation of the sentence in (2b) more often than in (2a).[8]

The ambiguity advantage in relative clauses and task-demands effects. The so-called ambiguity advantage is based on the observation of Traxler et al. (1998), who found a speedup at *mustache* in ambiguous conditions such as (3c) in comparison with unambiguous conditions such as (3a) and (3b), where *mustache* is the disambiguating word.

(3) a. The driver of the car that had the mustache was pretty cool. (high attachment)

b. The car of the driver that had the mustache was pretty cool. (low attachment)

c. The son of the driver that had the mustache was pretty cool. (globally ambiguous)

[8]Notice that even though it is unlikely that readers would understand that *the cabinets are in the box*, it may be that the sentence is reanalyzed when the parser reaches *box*.

The account of the activation-based model with different variances is very similar to the unrestricted race model proposed by van Gompel, Pickering, and Traxler (2000), which predicts statistical facilitation in the case of ambiguity. According to the unrestricted race model, the parser starts building all possible structures simultaneously. While the time taken depends on plausibility, it is also affected by noise. This means that the adopted structure in each trial is the one that takes the least time, leading to shorter time on average when there are more candidates.

The activation-based model with different variances would yield similar predictions to the unrestricted race model if the inhibitory effect of interference in (3c) is sufficiently small. Given the relatively small magnitude of interference effect in the literature (Jäger et al., submitted; Nicenboim et al., submitted), this is likely to be the case. In unambiguous cases such as (3a) and (3b), there is only one NP that matches the retrieval cue: "being capable of having a mustache" (i.e., *the driver*). In ambiguous cases such as (3c), there are two NPs that match the retrieval cue (i.e, *The son* and *the driver*). Therefore, we would expect statistical facilitation (similarly to the case of facilitatory interference) leading to faster reading times on average. This situation can be simulated using the *race-plot* script similarly as before: In (3a) or (3b), *the driver* (full match) can be assigned an activation of 5 (and $\sigma = 1$), and *the car* (partial match) can be assigned an activation of 2.5 (and $\sigma = 2$); this would result in a mean reading time of ≈416 ms. In (3c), both *The son* and *the driver* should have similar activation since there is no penalty component involved, both are a full match. However, the cue "can have a mustache" does not uniquely identify any candidate. Given the small magnitude of inhibitory interference effects, we could assume an activation of 4.8 (instead of 5) and the same variance since there is no mismatch (i.e. $\sigma = 1$) for both NPs. This would result in a faster reading time on average, namely ≈349 ms.

Furthermore, the activation-based model may be able to account for Logačev and Vasishth's (2016) observation that the parser seems to behave in a way that resembles a

race between interpretations (low attachment vs. high attachment) but it is also task-dependent (as assumed by Swets et al., 2008). This could be achieved by setting the timeout (the parameters of the accumulator associated with the retrieval failure) to be task-dependent: longer timeouts when instructions or context encourage attentive reading and shorter timeouts when a full interpretation is not needed for successfully completing the experimental task.

In this case, the direct access model could also predict the ambiguity advantage in a very simple way: While in (3a) or (3b) it is possible to retrieve the incorrect NP (i.e. "the car") leading to a certain proportion of slower backtracked retrievals, in (3c) there should only be fast directly accessed retrievals, since both NPs (i.e. "The son" and "the driver") are correct targets. In addition, for the direct access model, the proportion of incorrect retrievals that are backtracked could be task dependent, with a larger proportion of backtracking associated with deeper processing. However, the predictions of both models would not be identical. The direct access model predicts that the reading times at the disambiguating region when the incorrect interpretation (or no interpretation) is held in (3a) or (3b) would be identical to the reading times in (3c). In contrast, for the activation based-model with different variances, the relationship between reading times at the different conditions would depend on the comprehension accuracy. Future work that includes measures of reading times and queries for the comprehension of the relative clause, as well as manipulates task demands could compare the activation-based model with different variances, the direct access model, and the model presented in Logačev and Vasishth (2016), which subsumes the unrestricted race model and allows it to be task-dependent.

Good-enough processing. While a comprehensive alternative to good-enough processing is out of the scope of this section (see Christianson, 2016, for a complete overview), it should be noticed that without further assumptions the activation-based model with different variances and the direct access model can account for manipulations that show (sometimes unexpected) fast reading times which have been attributed to

good-enough processing. For the activation-based model with different variances, this can be achieved by associating the timeout with either task demands as suggested previously or also with individual differences. An increase of either timeout speed (i.e., the rate of accumulation of the retrieval failure) or an increase of its noise (i.e., the variance of the accumulator associated with retrieval failure) would lead to more frequent shallow parses with incomplete dependencies which are read faster. Thus, experiments that probe the comprehension of certain dependencies less often may lead to faster (and maybe noisier) timeouts, which would in turn lead to shorter mean reading times. Individual differences in participants such as working memory capacity may have a similar effect with lower capacity leading to faster and noisier timeouts in the retrieval process.

Similarly for the direct access model, the probability of backtracking could be affected by task demands or by individual differences: A less demanding task would reduce reading times and comprehension accuracy on average by discouraging backtracking. Individual differences may have a similar effect, participants with lower working memory capacity may be less prone to backtracking. As suggested before, this could be assessed in future work by including measures of reading times and comprehension accuracy of the relevant dependencies.

Conclusion

We have provided an evaluation of two theoretically grounded and empirically successful models in explaining similarity-based interference in sentence comprehension: Lewis and Vasishth's (2005) activation-based model built under the assumptions of ACT-R (Anderson et al., 2004) and McElree's (2000) direct access model. We also evaluated a third model, a variation of the activation-based model.

Our evaluation, which consisted in implementing these models in a Bayesian hierarchical framework, confirms that, as expected, both the activation-based and direct access models can in principle explain interference effects. However, posterior predictive

checks and model comparison using cross-validation show that some aspects of the data fit better under the direct access model, in particular, the default activation-based cannot predict that, on average, incorrect retrievals would be faster than correct ones.

However, we show that by introducing a modification of the activation model, namely, by assuming that the accumulation of evidence for the retrieval of incorrect items is not only slower but noisier (i.e., different variances for the correct and incorrect items), the new model can provide a fit as good as the one of the direct access model.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*(6), 716–723.

Anderson, J. R., Bothell, D., Lebiere, C., & Matessa, M. (1998). An integrated theory of list memory. *Journal of Memory and Language*, *38*(4), 341–380. doi:10.1006/jmla.1997.2553

Anderson, J. R., Bothell, D., Byrne, M. D., Douglass, S., Lebiere, C., & Qin, Y. (2004). An integrated theory of the mind. *Psychological Review*, *111*(4), 1036–1060. doi:10.1037/0033-295x.111.4.1036

Anderson, J. R. & Lebiere, C. (1998). *The atomic components of thought*. Mahwah: Erlbaum.

Anderson, J. R. & Reder, L. M. (1999). The fan effect: New results and new theories. *Journal of Experimental Psychology*, *128*(2), 186. doi:10.1037/0096-3445.128.2.186

Anderson, J. R., Reder, L. M., & Lebiere, C. (1996). Working memory: Activation limitations on retrieval. *Cognitive Psychology*, *30*(3), 221–256. doi:10.1006/cogp.1996.0007

Audley, R. J. & Pike, A. R. (1965). Some alternative stochastic models of choice. *British Journal of Mathematical and Statistical Psychology*, *18*(2), 207–225. doi:10.1111/j.2044-8317.1965.tb00342.x

Brown, S. D. & Heathcote, A. (2005). A ballistic model of choice response time. *Psychological Review*, *112*(1), 117–128. doi:10.1037/0033-295X.112.1.117

Brown, S. D. & Heathcote, A. (2008). The simplest complete model of choice response time: linear ballistic accumulation. *Cognitive psychology*, *57*(3), 153–178. doi:10.1016/j.cogpsych.2007.12.002

Chang, W., Cheng, J., Allaire, J., Xie, Y., & McPherson, J. (2016). *Shiny: web application framework for r*. R package version 0.14.

Christianson, K. (2016). When language comprehension goes wrong for the right reasons: good-enough, underspecified, or shallow language processing. *The Quarterly Journal of Experimental Psychology*, *69*(5), 817–828. doi:10.1080/17470218.2015.1134603

Cowan, N. (1995). *Attention and memory: An integrated framework*. Oxford University Press.

Daily, L. Z., Lovett, M. C., & Reder, L. M. (2001). Modeling individual differences in working memory performance: A source activation account. *Cognitive Science*, *25*, 315–353. doi:10.1016/S0364-0213(01)00039-8

Dillon, B., Chow, W. y., Wagers, M., Guo, T., Liu, F., & Phillips, C. (2014). The structure-sensitivity of memory access: Evidence from Mandarin Chinese. *Frontiers in Psychology*, *5*(1025). doi:10.3389/fpsyg.2014.01025

Dillon, B., Mishler, A., Sloggett, S., & Phillips, C. (2013). Contrasting intrusion profiles for agreement and anaphora: Experimental and modeling evidence. *Journal of Memory and Language*, *69*(2), 85–103. doi:10.1016/j.jml.2013.04.003

Engelmann, F. (2015). *Toward an integrated model of sentence processing in reading* (Doctoral dissertation, University of Potsdam).

Ferreira, F., Bailey, K. G. D., & Ferraro, V. (2002). Good-enough representations in language comprehension. *Current Directions in Psychological Science*, *11*(1), 11–15. doi:10.1111/1467-8721.00158

Foraker, S. & McElree, B. (2011). Comprehension of linguistic dependencies: Speed–accuracy tradeoff evidence for direct–access retrieval from memory. *Language and Linguistics Compass*, *5*(11), 764–783. doi:10.1111/j.1749-818X.2011.00313.x.Comprehension

Geisser, S. & Eddy, W. F. (1979). A predictive approach to model selection. *Journal of the American Statistical Association*, *74*(365), 153–160. doi:10.2307/2286745

Gelman, A., Carlin, J., Stern, H., Dunson, D., Vehtari, A., & Rubin, D. (2014). Bayesian data analysis. (Third). Taylor & Francis.

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

Gelman, A., Hill, J., & Yajima, M. (2012). Why we (usually) don't have to worry about multiple comparisons. *Journal of Research on Educational Effectiveness*, *5*(2), 189–211. doi:10.1080/19345747.2011.618213

Gelman, A., Hwang, J., & Vehtari, A. (2014). Understanding predictive information criteria for Bayesian models. *Statistics and Computing*, *24*(6), 997–1016. doi:10.1007/s11222-013-9416-2

Gibson, E. (2000). The dependency locality theory: A distance-based theory of linguistic complexity. In *Image, language, brain: Papers from the first mind articulation project symposium* (pp. 95–126).

Gillund, G. & Shiffrin, R. M. (1984). A retrieval model for both recognition and recall. *Psychological Review*, *91*(1), 1. doi:10.1037/0033-295X.91.1.1

Gordon, P. C., Hendrick, R., & Levine, W. H. (2002). Memory-load interference in syntactic processing. *Psychological Science*, *13*(5), 425–430. doi:10.1111/1467-9280.00475

Grodner, D. & Gibson, E. (2005). Consequences of the serial nature of linguistic input for sentenial complexity. *Cognitive Science*, *29*(2), 261–290. doi:10.1207/s15516709cog0000_7

Heathcote, A. & Love, J. (2012). Linear deterministic accumulator models of simple choice. *Frontiers in Psychology*, *3*. doi:10.3389/fpsyg.2012.00292

Hintze, J. L. & Nelson, R. D. (1998). Violin plots: A box plot-density trace synergism. *The American Statistician*, *52*(2), 181–184. doi:10.1080/00031305.1998.10480559

Jäger, L. A., Engelmann, F., & Vasishth, S. (2015). Retrieval interference in reflexive processing: Experimental evidence from Mandarin, and computational modeling. *Frontiers in Psychology*, *6*(617). doi:10.3389/fpsyg.2015.00617

Jäger, L. A., Engelmann, F., & Vasishth, S. (submitted). *Similarity-based interference in sentence comprehension: Literature review and Bayesian meta-analysis*. Manuscript submitted.

Kleinschmidt, D., Fine, A. B., & Jaeger, T. F. (2012). A belief-updating model of adaptation and cue combination in syntactic comprehension. In *Proceedings of the 34rd Annual Meeting of the Cognitive Science Society (CogSci12)* (pp. 605–10).

Kush, D. & Phillips, C. (2014). Local anaphor licensing in an SOV language: Implications for retrieval strategies. *Frontiers in Psychology*, *5*(1252). doi:10.3389/fpsyg.2014.01252

Lebiere, C. (1999). The dynamics of cognition: An ACT-R model of cognitive arithmetic. *Kognitionswissenschaft*, *8*(1), 5–19. doi:10.1007/BF03354932

Lebiere, C., Anderson, J. R., & Reder, L. M. (1994). Error modeling in the ACT-R production system. In *Proceedings of the Sixteenth Annual Conference of the Cognitive Science Society* (pp. 555–559). Erlbaum Hillsdale, NJ.

Lee, M. D. (2011). How cognitive modeling can benefit from hierarchical Bayesian models. *Journal of Mathematical Psychology*, *55*(1), 1–7. doi:10.1016/j.jmp.2010.08.013

Lewandowski, D., Kurowicka, D., & Joe, H. (2009). Generating random correlation matrices based on vines and extended onion method. *Journal of multivariate analysis*, *100*(9), 1989–2001. doi:10.1016/j.jmva.2009.04.008

Lewis, R. L. & Vasishth, S. (2005). An activation-based model of sentence processing as skilled memory retrieval. *Cognitive Science*, *29*(3), 375–419. doi:10.1207/s15516709cog0000_25

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

Logačev, P. & Vasishth, S. (2016). A multiple-channel model of task-dependent ambiguity resolution in sentence comprehension. *Cognitive Science*, *40*(2), 266–298. doi:10.1111/cogs.12228

Marcus, G. F. (2013). Evolution, Memory, and the Nature of Syntactic Representation. In Johan J. Bolhuis & M. Everaert (Eds.), *Birdsong, speech, and language: Exploring the evolution of mind and brain* (Chap. 2, pp. 27–44). MIT press.

McElree, B. (1993). The locus of lexical preference effects in sentence comprehension: A time-course analysis. *Journal of Memory and Language*, *32*(4), 536–571. doi:10.1006/jmla.1993.1028

McElree, B. (2000). Sentence comprehension is mediated by content-addressable memory structures. *Journal of Psycholinguistic Research*, *29*(2), 111–123. doi:10.1023/A:1005184709695

McElree, B. (2006). Accessing recent events. *Psychology of Learning and Motivation Volume 46*, 155–200. doi:10.1016/S0079-7421(06)46005-9

McElree, B. & Dosher, B. A. (1989). Serial position and set size in short-term memory: The time course of recognition. *Journal of Experimental Psychology: General*, *118*(4), 346–373. doi:10.1037/0096-3445.118.4.346

McElree, B., Foraker, S., & Dyer, L. (2003). Memory structures that subserve sentence comprehension. *Journal of Memory and Language*, *48*(1), 67–91. doi:10.1016/s0749-596x(02)00515-6

Myslín, M. & Levy, R. (2016). Comprehension priming as rational expectation for repetition: Evidence from syntactic processing. *Cognition*, *147*, 29–56. doi:10.1016/j.cognition.2015.10.021

Nairne, J. S. (1990). A feature model of immediate memory. *Memory & Cognition*, *18*(3), 251–269. doi:10.3758/BF03213879

Nicenboim, B., Engelmann, F., Suckow, K., & Vasishth, S. (submitted). Number interference in german: Implications for theories of cue-based retrieval.

Nicenboim, B., Logačev, P., Gattei, C., & Vasishth, S. (2016). When high-capacity readers slow down and low-capacity readers speed up: Working memory and locality effects. *Frontiers in Psychology*, *7*(280). doi:10.3389/fpsyg.2016.00280

Nicenboim, B. & Vasishth, S. (2016). Statistical methods for linguistic research: Foundational ideas - Part II. *Language and Linguistics Compass*. In Press.

Nicenboim, B., Vasishth, S., Gattei, C., Sigman, M., & Kliegl, R. (2015). Working memory differences in long-distance dependency resolution. *Frontiers in Psychology*, *1*. doi:10.3389/fpsyg.2015.00312

Oberauer, K. (2013). The focus of attention in working memory—from metaphors to mechanisms. *Frontiers in Human Neuroscience*, *7*, 673. doi:10.3389/fnhum.2013.00673

Oberauer, K. & Kliegl, R. (2006). A formal model of capacity limits in working memory. *Journal of Memory and Language*, *55*(4), 601–626. doi:10.1016/j.jml.2006.08.009

Papaspiliopoulos, O., Roberts, G. O., & Sköld, M. (2007). A general framework for the parametrization of hierarchical models. *Statistical Science*, *22*(1), 59–73. doi:10.1214/088342307000000014

Parker, D. & Phillips, C. (201629). Negative polarity illusions and the format of hierarchical encodings in memory. *Cognition*. in press.

R Core Team. (2015). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria.

Raab, D. H. (1962). Division of psychology: Statistical facilitation of simple reaction times. *Transactions of the New York Academy of Sciences*, *24*(5 Series II), 574–590.

Raaijmakers, J. G. W. & Shiffrin, R. M. (1980). SAM: A theory of probabilistic search of associative memory. *The psychology of Learning and Motivation*, *14*, 207–262.

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, *85*(2), 59. doi:10.1037/0033-295X.85.2.59

Ratcliff, R. & McKoon, G. (2008). The diffusion decision model: Theory and data for two-choice decision tasks. *Neural Computation*, *20*(4), 873–922. doi:10.1162/neco.2008.12-06-420

Ratcliff, R. & Rouder, J. N. (1998). Modeling Response Times for Two-Choice Decisions. *Psychological Science*, *9*(5), 347–356. doi:10.1111/1467-9280.00067

Ratcliff, R., Smith, P. L., Brown, S. D., & McKoon, G. (2016). Diffusion decision model: Current issues and history. *Trends in Cognitive Sciences*, *20*(4), 260–281. doi:10.1016/j.tics.2016.01.007

Reed, A. V. (1973). Speed-accuracy trade-off in recognition memory. *Science*, *181*(4099), 574–576. doi:10.1126/science.181.4099.574

Rouder, J. N. (2005). Are unshifted distributional models appropriate for response time? *Psychometrika*, *70*(2), 377–381. doi:10.1007/s11336-005-1297-7

Rouder, J. N., Province, J. M., Morey, R. D., Gomez, P., & Heathcote, A. (2014). The lognormal race: A cognitive-process model of choice and latency with desirable psychometric properties. *Psychometrika*, *80*(2), 491–513. doi:10.1007/s11336-013-9396-3

Rouder, J. N., Tuerlinckx, F., Speckman, P. L., Lu, J., & Gomez, P. (2008). A hierarchical approach for fitting curves to response time measurements. *Psychonomic Bulletin & Review*, *15*(6), 1201–1208. doi:10.3758/pbr.15.6.1201

Shiffrin, R. M., Lee, M., Kim, W., & Wagenmakers, E.-J. (2008). A survey of model evaluation approaches with a tutorial on hierarchical Bayesian methods. *Cognitive Science*, *32*(8), 1248–1284. doi:10.1080/03640210802414826

Sorensen, T., Hohenstein, S., & Vasishth, S. (2016). Bayesian linear mixed models using Stan: A tutorial for psychologists, linguists, and cognitive scientists. *Quantitative Methods for Psychology*. In Press.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *64*(4), 583–639. doi:10.1111/1467-9868.00353

Stan Development Team. (2016a). Rstan: the R interface to Stan, version 2.9.0.

Stan Development Team. (2016b). Stan: A C++ library for probability and sampling, version 2.9.0.

Stan Development Team. (2016c). *Stan modeling language users guide and reference manual, version 2.9*.

Summerfield, C. & Tsetsos, K. (2015). Do humans make good decisions? *Trends in Cognitive Sciences*, *19*(1), 27–34. doi:10.1016/j.tics.2014.11.005

Swets, B., Desmet, T., Clifton, C. J., & Ferreira, F. (2008). Underspecification of syntactic ambiguities: evidence from self-paced reading. *Memory & Cognition*, *36*(1), 201–216. doi:10.3758/MC.36.1.201

Traxler, M. J. (2014). Trends in syntactic parsing: Anticipation, Bayesian estimation, and good-enough parsing. *Trends in Cognitive Sciences*, *18*(11), 605–611. doi:10.1016/j.tics.2014.08.001

Traxler, M. J., Pickering, M. J., & Clifton, C. J. (1998). Adjunct attachment is not a form of lexical ambiguity resolution. *Journal of Memory and Language*, *39*(4), 558–592. doi:http://dx.doi.org/10.1006/jmla.1998.2600

Usher, M. & McClelland, J. L. (2001). The time course of perceptual choice: The leaky, competing accumulator model. *Psychological Review*, *108*(3), 550. doi:10.1037/0033-295X.108.3.550

Van Dyke, J. A. (2007). Interference effects from grammatically unavailable constituents during sentence processing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *33*(2), 407. doi:10.1037/0278-7393.33.2.407

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*(3), 285–316. doi:10.1016/S0749-596X(03)00081-0

Van Dyke, J. A. & McElree, B. (2006). Retrieval interference in sentence comprehension. *Journal of Memory and Language*, *55*(2), 157–166. doi:10.1016/j.jml.2006.03.007

Van Dyke, J. A. & McElree, B. (2011). Cue-dependent interference in comprehension. *Journal of Memory and Language*, *65*(3), 247–263. doi:10.1016/j.jml.2011.05.002

van Gompel, R. P. G., Pickering, M. J., & Traxler, M. J. (2000). Unrestricted race: A new model of syntactic ambiguity resolution. *Reading as a perceptual process*, 621–648. doi:/10.1016/B978-008043642-5/50029-2

Van Rijn, H. & Anderson, J. R. (2003). Modeling lexical decision as ordinary retrieval. In *In Proceedings of the International Conference on Cognitive Modeling (ICCM)* (pp. 207–212).

van Maanen, L., van Rijn, H., & Taatgen, N. (2011). RACE/A: An architectural account of the interactions between learning, task control, and retrieval dynamics. *Cognitive Science*, *36*(1), 62–101. doi:10.1111/j.1551-6709.2011.01213.x

Vasishth, S., Brüssow, S., Lewis, R. L., & Drenhaus, H. (2008). Processing polarity: How the ungrammatical intrudes on the grammatical. *Cognitive Science*, *32*(4), 685–712. doi:10.1080/03640210802066865

Vasishth, S. & Lewis, R. L. (2006). Argument-head distance and processing complexity: Explaining both locality and antilocality effects. *Language*, *82*(4), 767–794. doi:10.1353/lan.2006.0236

Vehtari, A. & Gelman, A. (2015). Pareto smoothed importance sampling. *arXiv preprint arXiv:1507.02646*.

Vehtari, A., Gelman, A., & Gabry, J. (2015). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *ArXiv e-prints*. arXiv: 1507.04544 [stat.CO]

Vehtari, A., Gelman, A., & Gabry, J. (2016). *Loo: Efficient leave-one-out cross-validation and WAIC for Bayesian models*. R package version 0.1.6.

Vehtari, A. & Ojanen, J. (2012). A survey of Bayesian predictive methods for model assessment, selection and comparison. *Statistics Surveys*, *6*, 142–228. doi:10.1214/12-ss102

Vickers, D. (1970). Evidence for an accumulator model of psychophysical discrimination. *Ergonomics*, *13*(1), 37–58. doi:10.1080/00140137008931117

Wagenmakers, E.-J., Ratcliff, R., Gomez, P., & McKoon, G. (2008). A diffusion model account of criterion shifts in the lexical decision task. *Journal of Memory and Language*, *58*(1), 140–159. doi:10.1016/j.jml.2007.04.006

Wagers, M. W., Lau, E. F., & Phillips, C. (2009). Agreement attraction in comprehension: representations and processes. *Journal of Memory and Language*, *61*(2), 206–237. doi:10.1016/j.jml.2009.04.002

Wickelgren, W. A. (1977). Speed-accuracy tradeoff and information processing dynamics. *Acta psychologica*, *41*(1), 67–85. doi:10.1016/0001-6918(77)90012-9

Xu, F. & Tenenbaum, J. B. (2007). Sensitivity to sampling in Bayesian word learning. *Developmental Science*, *10*(3), 288–297. doi:10.1111/j.1467-7687.2007.00590.x

Zandbelt, B., Purcell, B. A., Palmeri, T. J., Logan, G. D., & Schall, J. D. (2014). Response times from ensembles of accumulators. *Proceedings of the National Academy of Sciences*, *111*(7), 2848–2853. doi:10.1073/pnas.1310577111

Appendix A

Implementation of the activation-based model in Stan

The Stan code (shown in Listing 1) was fit to a Latin-squared design, where only the sentences of the original experiment (Nicenboim et al., submitted) with questions that queried the subject of the embedded verb was kept, and it used a non-centered parameterization to improve convergence (for details see: Papaspiliopoulos, Roberts, & Sköld, 2007; Stan Development Team, 2016c) in Stan (Stan Development Team, 2016b). However, to improve clarity, we ignore that each participant did not respond to each experimental item, and we assume a centered parametrization in the equations below.

Let $i = 1, ..., N_{subj}$, $j = 1, ..., N_{items}$, and $c = 1, ..., N_{choices}$ index participants, items, and choices in the multiple-choice questions (1 is the correct response, the target of the retrieval, 2 and 3 are incorrect responses, the competitors, and 4 is the option "I don't know", which represents a failed retrieval) respectively. Let $w_{i,j}$, and $RT_{i,j}$ denote the response selected and the reading times at the auxiliary verb (*hatte*) for subject i to the item j. Then we assume that reading times have the following distribution:

$$RT_{i,j} \sim \psi_i + lognormal(b - \alpha_{i,j,c=w}, \sigma) \tag{45}$$

where ψ_i is a by-subject shift, b is an arbitrary threshold (set to 10), and $\alpha_{i,j,c=w}$ represents the rate of accumulation of the "winner" accumulator. The rest of the accumulators that did not win the race must have been slower in that specific trial. From this it follows that the accumulators that lost the race have a potential $RT'_{i,j,c\neq w}$ which is larger than the observed value $RT_{i,j}$.

If all the answers are selected at least once (and if not, we can safely remove the accumulator since its rate of accumulation is so low that it never wins), the race turns into a problem of censored data, where the reading times, $RT'_{i,j,c\neq w}$, above an upper bound, $RT_{i,j,c=w}$, never occur. In order to calculate the posterior of the rate of accumulation, α, of all the accumulators, we cannot ignore the censored data (Gelman, Carlin, et al., 2014, pp.

224-227). However, it is not necessary to impute values, and the values can be integrated out (Stan Development Team, 2016c, pp. 107–110; Gelman, Carlin, et al., 2014, pp. 224-227). Each censored data point has a probability of

$$Pr[RT'_{i,j,c\neq w} > RT_{i,j}] = \int_{RT_{i,j}-\psi_i}^{\infty} lognormal(RT'_{i,j,c\neq w} - \psi_i | \alpha_{i,j,c\neq w}, \sigma) \tag{46}$$

$$= 1 - \Phi\left(\frac{log(RT_{i,j} - \psi_i) - \alpha_{i,j,c\neq w}}{\sigma}\right) \tag{47}$$

where $\Phi()$ is the cumulative distribution function of the standard normal distribution. Since the shifts of the distribution, ψ_i, must be positive, to ensure convergence of the model we exponentiate a term that is associated with a general shift of the whole reading times distribution, ψ', and a term that represents the by-participants adjustment, ψ'_i:

$$\psi_i = exp(\psi' + \psi'_i) \tag{48}$$

with the following priors for the by-participant component:

$$\psi'_i \sim normal(0, \tau_\psi) \tag{49}$$

$$\tau_\psi \sim normal(0, 0.5) \tag{50}$$

In addition, each ψ_i must be smaller than the shortest reading time of each participant i (recall that the shift is the lower bound of the distribution). We satisfied the constraint on the upper bound with the following prior on the general shift:

$$\psi' \sim normal(0, log(mean(RT)))T[, U] \tag{51}$$

where a normal distribution is truncated on the upper limit, U, which is the smallest difference between $log(RT)$ and ψ_i'.

We assume that the rates of accumulation depend on the experimental condition (high or low interference) and that the rates may be affected by participants and by items. We can express this in matrix notation for each accumulator as follows:

$$\alpha_c = \mathbf{X}\boldsymbol{\beta}_c + \mathbf{X}\mathbf{u}_c + \mathbf{X}\mathbf{v}_c \tag{52}$$

Here $\mathbf{X}$ is the $N_{obs} \times N_{pars}$ model matrix, with the number of parameters (so-called fixed effects), N_{pars}, being two: intercept and condition. Each $\boldsymbol{\beta}_c$ is a vector of length N_{pars} with the estimates of the fixed-effect parameters for the accumulator associated with the choice c. Each $\mathbf{u}_c$ and $\mathbf{v}_c$ are the by-participants and by-item adjustments to the fixed effects estimates (so-called random-effects) for the accumulator c. We used weakly informative priors for all the estimates (some estimates were reparametrized in the Stan implementation, see Listing 1 for details). The priors for the fixed effects were set as follows:

$$\beta_{0,c} \sim normal(0, 10 - log(mean(RT_c))) \tag{53}$$

$$\beta_{1,c} \sim normal(0, 1) \tag{54}$$

Here, $\beta_{0,c}$ are the intercepts of the fixed effects for choice c, $log(mean(RT_c))$ is the logarithm of the mean of the reading times when option c was selected, and $\beta_{1,c}$ represents the slopes of the fixed effects (i.e., the effect of interference).

All the random-effects, $\mathbf{u}_c$, and $\mathbf{v}_c$, were assumed to be sampled from two multivariate normal distributions with means of zero. The prior of the standard deviations of the random effects was $normal(0, 1)$. We placed lkj priors on the random effects correlation matrices with shape parameter $\eta = 2$ (see Lewandowski, Kurowicka, & Joe,

2009; Sorensen, Hohenstein, & Vasishth, 2016).

Listing 1: Stan code for the activation-based model

```
functions {
  real shift_max(vector shift_u, int[] subj, vector rt, real logmeanrt) {
    real shift_max;
    shift_max <- positive_infinity();
    for (i in 1:num_elements(rt)){
      shift_max <- fmin(shift_max,log(rt[i]) - shift_u[subj[i]] );
    }
    return (shift_max/logmeanrt);
  }

  real race_ACTR_log(int winner, real shifted_rt, row_vector activation, real
     threshold, real noise){
    real log_lik;
    int N_choices;
    N_choices <- cols(activation);
    log_lik <- 0;
    for(l in 1:N_choices){
      if(l == winner){
        log_lik <- log_lik + lognormal_log(shifted_rt,threshold-activation[l],
           noise);
      } else {
        log_lik <- log_lik + lognormal_ccdf_log
                 (shifted_rt,threshold-activation[l], noise);
      }
    }
    return(log_lik);
  }

}

data {
  int<lower=0> N_obs;
  vector<lower=0>[N_obs] rt;
  int N_choices;
  int<lower=1> winner[N_obs];
  int N_coef;
  matrix[N_obs,N_coef] x;
  int<lower=0> N_coef_u;
  int<lower=1> subj[N_obs];
  int<lower=1> N_subj;
  matrix[N_obs,N_coef_u] x_u;
  int<lower=0> N_coef_w;
  int<lower=1> item[N_obs];
  int<lower=1> N_item;
  matrix[N_obs,N_coef_w] x_w;
}

transformed data {
  matrix[N_obs,N_coef-1] x_betas;
```

```
  int N_tau_u;
  int N_tau_w;
  real logmeanrt;
  vector[N_choices] logmeanrtw; #mean by winner
  vector[N_choices] mean_rtw; #mean by winner
  real scaling;

  logmeanrt <- log(mean(rt));
  x_betas <- x[,2:N_coef];
  N_tau_u <- N_coef_u * N_choices + 1;
  N_tau_w <- N_coef_w * N_choices + 1;
  {
    matrix[N_obs,N_choices] winner1;
    for(i in 1:N_obs){
      for(j in 1:N_choices){
        winner1[i,j] <-  (j ==  winner[i]);
      }
    }
    mean_rtw <- ((rt' * winner1) /  //sum of rts per winner
    crossprod(winner1))';  # x^T * x (length of winners)
    logmeanrtw <- log(mean_rtw);
  }
  scaling <- 10;
}

parameters{
  real<lower=0> sigma;
  real alpha_raw[N_choices];
  vector<lower=0> [N_tau_u]  tau_u;     // subj sd
  cholesky_factor_corr[N_tau_u] L_u;    // corr. matrix for random intercepts
     and slopes by subj
  matrix[N_tau_u,N_subj] z_u;
  vector<lower=0> [N_tau_w]  tau_w;     // item sd
  cholesky_factor_corr[N_tau_w] L_w;    // corr. matrix for random intercepts
     and slopes by items
  matrix[N_tau_w,N_item] z_w;
  real<lower=0> tau_shift;     // by subj sd of shift
  vector[N_subj] shift_u_raw;  // by subj scaled shift
  real<upper=shift_max(shift_u_raw * tau_shift, subj, rt, logmeanrt)>
     shift_raw;
  matrix[N_coef-1,N_choices] beta;
}

transformed parameters {
  matrix[N_coef_u,N_choices]  u[N_subj];
  matrix[N_coef_w,N_choices]  w[N_item];
  vector[N_subj] shift_u;  // by subj shift
  real<lower=0> shift;
  row_vector[N_choices] alpha;

  {
    matrix[N_tau_u,N_subj] u_long;
    matrix[N_tau_w,N_item] w_long;
    matrix[N_tau_u,N_tau_u] Lambda_u;
```

```
    matrix[N_tau_w,N_tau_w] Lambda_w;
    Lambda_u <- diag_pre_multiply(tau_u,L_u);
    Lambda_w <- diag_pre_multiply(tau_w,L_w);
    u_long <- (Lambda_u * z_u);
    w_long <- (Lambda_w * z_w);
    for (i in 1:N_subj){
      for (j in 1:N_choices){
        u[i, , j] <- u_long[(j - 1) * N_coef_u + 1 : j * N_coef_u, i];
      }
    }

    for (i in 1:N_item){
      for (j in 1:N_choices){
        w[i, , j] <- w_long[(j - 1) * N_coef_w + 1 : j * N_coef_w, i];
      }
    }
  }

  shift_u <- shift_u_raw * tau_shift; // = shift_u ~ normal(0,tau_shift)
  shift <- shift_raw * logmeanrt;

  for (j in 1:(N_choices))
    alpha[j] <- scaling - alpha_raw[j] .* logmeanrtw[j];
}
model {
  sigma ~ normal(0,2);
  alpha_raw ~ normal(0,1);
  to_vector(beta) ~ normal(0,1);
  to_vector(z_u) ~ normal(0,1);
  to_vector(z_w) ~ normal(0,1);
  tau_u ~ normal(0,1);
  tau_w ~ normal(0,1);
  L_u ~ lkj_corr_cholesky(2.0);
  L_w ~ lkj_corr_cholesky(2.0);
  tau_shift  ~ normal(0,.5);
  shift_u_raw ~ normal(0,1);
  shift_raw ~ normal(0,1);

  for ( i in 1 : N_obs ) {
    row_vector[N_choices] A;
    real shifted_rt;
    shifted_rt <- rt[i] - exp(shift + shift_u[subj[i]]);

    A[1:N_choices] <- alpha[1:N_choices]
      + x_betas[i] * beta  +
        x_u[i] * u[subj[i],]  +
        x_w[i] * w[item[i],];

    winner[i] ~ race_ACTR(shifted_rt,A, scaling,sigma);
  }
}
```

Appendix B
Implementation of the direct access model in Stan

The code (shown in Listing 2) was fit to the same data as the activation-based model. As before, to improve clarity, we ignore that each subject did not respond to each item and we assume a centered parametrization.

Let $i = 1, ..., N_{subj}$, $j = 1, ..., N_{items}$, and $c = 1, ..., N_{choices}$ index participants, items, and choices respectively, where choice 1 is the correct response and choice $N_{choices}$ (which maps to 4) is the response associated with a retrieval failure. Let $w_{i,j}$, and $RT_{i,j}$ denote the response selected and the reading times at the auxiliary verb (*hatte*) for subject i to the item j.

We implemented the assumptions of the direct access model, by letting w have a discrete distribution that follows a one-inflated categorical model, where additional probability mass is added to the outcome 1 (correct response) due to backtracking with probability P_b as follows:

$$P(w_{i,j} = 1|\boldsymbol{\theta_{i,j}}, P_b) = Categorical(y = 1|\boldsymbol{\theta_{i,j}}) + (1 - Categorical(y = 1|\boldsymbol{\theta_{i,j}})) \cdot P_b \quad (55)$$

$$P(w_{i,j} = s|\boldsymbol{\theta_{i,j}}, P_b) = Categorical(y = s|\boldsymbol{\theta_{i,j}}) \cdot (1 - P_b), \text{ with } s > 1 \quad (56)$$

where $\boldsymbol{\theta}$ is a vector of $N_{choices}$ rows that represents the probability of each option.

If the answer given is wrong, we assume that there is no backtracking and then reading times are distributed in the following way:

$$RT_{i,j,\forall w,w>1} \sim \psi_i + lognormal(T_{da,i,j}, \sigma) \quad (57)$$

where ψ_i is a by-subject shift, T_{da} represents the time needed for the direct access or failure together with extra processes

If the answer given is right, reading times are assumed to have a mixture distribution. This is so because there are two "paths" to reach a correct response (see Figure 5): (i) The chunk that is retrieved is the correct one (at the first try), and this means that there is direct access and reading times should belong to a distribution similar to the previous one as shown in Equation (57); or (ii) an incorrect chunk (or no chunk) is retrieved but is backtracked, and this means that reading times should belong to a distribution with a larger location than $T_{da,i,j}$, namely, $T_{da,i,j} + t_{b,i,j}$. Thus RTs should be distributed in the following way:

$$RT_{i,j,w=1} \sim \psi_i \quad (58)$$

$$+ P(Categorical(y = 1|\boldsymbol{\theta_{i,j}})|w_{i,j} = 1) \cdot lognormal(T_{da,i,j}, \sigma) \quad (59)$$

$$+ P(Categorical(y \neq 1|\boldsymbol{\theta_{i,j}}) \cdot P_b|w_{i,j} = 1) \cdot lognormal(T_{da,i,j} + t_{b,i,j}, \sigma) \quad (60)$$

where, from Equation (55), $P(Categorical(y = 1|\boldsymbol{\theta_{i,j}})|w_{i,j} = 1)$ in (59) is

$$\frac{Categorical(y = 1|\boldsymbol{\theta_{i,j}})}{Categorical(y = 1|\boldsymbol{\theta_{i,j}}) + (1 - Categorical(y = 1|\boldsymbol{\theta_{i,j}})) \cdot P_b} \quad (61)$$

and $P(Categorical(y \neq 1|\boldsymbol{\theta_{i,j}}) \cdot P_b|w_{i,j} = 1)$ in Equation (60) is

$$\frac{(1 - Categorical(y = 1|\boldsymbol{\theta_{i,j}})) \cdot P_b}{Categorical(y = 1|\boldsymbol{\theta_{i,j}}) + (1 - Categorical(y = 1|\boldsymbol{\theta_{i,j}})) \cdot P_b} \tag{62}$$

Furthermore, the categorical model was fit including a hierarchical structure in $\boldsymbol{\theta'}$, where $\boldsymbol{\theta'}$ is a vector with $N_{choices}$ rows with its last row set to zero, so that $softmax(\boldsymbol{\theta'}) = \boldsymbol{\theta}$.[9] This way we ensure the identifiability of $Categorical(softmax(\boldsymbol{\theta'}))$.

We assume that the probability of each choice depend on the experimental condition (high or low interference) and that the probabilities may be affected by participants and by items. In matrix notation, the first $N_{choices} - 1$ rows of $\boldsymbol{\theta'}$ are structured as the activations in the activation-based model:

$$\theta_c = \mathbf{X}\boldsymbol{\beta}_c + \mathbf{X}\mathbf{u}_c + \mathbf{X}\mathbf{v}_c \tag{63}$$

As for the activation-based model, we used weakly informative priors for all the estimates. The priors for the fixed effects were set with the added constraint that $\beta_{1,0}$, the intercept of the probability of the correct choice (the first choice) in logit-space, was constrained to be larger than $\beta_{2..3,0}$, the intercept associated with the incorrect responses, and zero (which is the value associated with the last choice):

$$\beta_{2..3,c} \sim normal(0, 1) \tag{64}$$
$$\beta_{1,0} \sim normal(2, 2) + max(\beta_{2..3,c}, 0) \tag{65}$$
$$P_b \sim beta(1, 1) \tag{66}$$

In addition, we assumed a hierarchical structure to the parameters associated with latencies:

$$T_{da,i,j} = \beta_{Tda} + u_{Tda,i} + v_{Tda,j} \tag{67}$$
$$T_{b,i,j} = \beta_{Tb} + u_{Tb,i} + v_{Tb,j} \tag{68}$$

with the following priors on the intercepts:

$$\beta_{Tda} \sim normal(0, log(mean(RT))) \tag{69}$$
$$\beta_{Tb} \sim normal(0, 1) \tag{70}$$

All the random-effects, $\mathbf{u}_c$, u_{tda}, u_{tb}, $\mathbf{v}_c$, v_{tda}, v_{tb} were assumed to be sampled from four multivariate normal distributions with means of zero: (i) for the subject adjustment on probabilities of retrieval, (ii) for a similar adjustment for items, (iii) for the subject adjustment on latencies, and (iv) for a similar adjustment for items. As before we placed lkj priors on the random effects correlation matrices with shape parameter $\eta = 2$.

[9] The softmax function is defined as in Stan Development Team (2016c) by $softmax(y) = \frac{exp(y)}{\sum_{k=1}^{K} exp(y_k)}$

Listing 2: Stan code for the direct access model

```
functions {
  real shift_max(vector shift_u, int[] subj, vector rt, real logmeanrt) {
    real shift_max;
    shift_max <- positive_infinity();
    for (i in 1:num_elements(rt)){
      shift_max <- fmin(shift_max,log(rt[i]) - shift_u[subj[i]] );
    }
    return (shift_max/logmeanrt);
  }

 real da_log(int winner, real rt, real P_redo, vector mu_c, real mu_rt, real
              mu_rt_redo, real sigma){
  real logP_w1;

  logP_w1 <- log_sum_exp(categorical_logit_log(1,mu_c),
                         log(P_redo) + log1m_exp(categorical_logit_log(1,mu_c))
                            );
  if(winner==1) {
    return logP_w1 + log_sum_exp(log(P_redo) +
                      log1m_exp(categorical_logit_log(1,mu_c)) - logP_w1 +
                      lognormal_log(rt,mu_rt_redo,sigma),
                      categorical_logit_log(1,mu_c) - logP_w1 +
                      lognormal_log(rt,mu_rt,sigma));
    } else {
      return log1m(P_redo) + categorical_logit_log(winner,mu_c) +
             lognormal_log(rt,mu_rt,sigma);
    }
  }
}

data {
  int<lower=0> N_obs;
  vector<lower=0>[N_obs] rt;
  int N_choices;
  int<lower=1> winner[N_obs];
  int N_coef;
  matrix[N_obs,N_coef] x;
  int<lower=0> N_coef_u;
  int<lower=1> subj[N_obs];
  int<lower=1> N_subj;
  matrix[N_obs,N_coef_u] x_u;
  int<lower=0> N_coef_w;
  int<lower=1> item[N_obs];
  int<lower=1> N_item;
  matrix[N_obs,N_coef_w] x_w;
}

transformed data {
  matrix[N_obs,N_coef-1] x_betas;
  matrix[N_coef_u,N_obs] x_ut;
  matrix[N_coef_w,N_obs] x_wt;
  int N_tau_u;
```

```
52    int N_tau_w;
53    real logmeanrt;
54    vector[N_choices] logmeanrtw; #mean by winner
55    vector[N_choices] mean_rtw; #mean by winner
56    real scaling;
57
58    logmeanrt <- log(mean(rt));
59    x_betas <- x[,2:N_coef];
60    x_ut <- x_u';
61    x_wt <- x_w';
62    N_tau_u <- N_coef_u*(N_choices-1) ;
63    N_tau_w <- N_coef_w*(N_choices-1) ;
64    {
65      matrix[N_obs,N_choices] winner1;
66      for(i in 1:N_obs){
67        for(j in 1:N_choices){
68          winner1[i,j] <-  (j ==  winner[i]);
69        }
70      }
71      mean_rtw <- ((rt' * winner1) /  #sum of rts per winner
72      crossprod(winner1))';  # x^T * x (length of winners)
73      logmeanrtw <- log(mean_rtw);
74    }
75    scaling <- 10;
76  }
77
78  parameters{
79    real<lower=0,upper=1> P_redo;  // Prob. of backtracking
80    real<lower=0> mu_add;
81    real<lower=0> sigma;
82    matrix[N_choices-1,N_coef-1] beta;  #failure doesn't have betas
83    vector<lower=0>[N_tau_u+2] tau_u;  // subj sd
84    cholesky_factor_corr[N_tau_u] L_u; // corr. matrix for r.e. of prob. by subj
85    cholesky_factor_corr[2] L_rt_u;    // corr. matrix for r.e. latencies. by
          subj
86    matrix[N_tau_u,N_subj] z_u;
87    matrix[2,N_subj] z_rt_u;
88    vector<lower=0>[N_tau_w+2] tau_w;  // item sd
89    cholesky_factor_corr[N_tau_w] L_w; // corr. matrix for r.e. of prob. by item
90    cholesky_factor_corr[2] L_rt_w;    // corr. matrix for r.e. latencies. by
          item
91    matrix[N_tau_w,N_item] z_w;
92    matrix[2,N_item] z_rt_w;
93    real<lower=0> alpha_rt_raw;
94    vector[N_choices-2] alpha_c_wrong;
95    real<lower=0> add_c;
96    real<lower=0> tau_shift;     // by subj sd of shift
97    vector[N_subj] shift_u_raw;  // subj shift
98    real<upper=shift_max(shift_u_raw * tau_shift, subj, rt, logmeanrt)>
          shift_raw;
99  }
100
101 transformed parameters{
102   real<lower=0> alpha_rt;
```

```
  vector[N_choices-1] alpha_c;
  matrix[N_choices-1,N_coef_u] u[N_subj];
  matrix[N_choices-1,N_coef_w] w[N_item];
  real u_rt[N_subj];
  real u_rt_redo[N_subj];
  real w_rt[N_item];
  real w_rt_redo[N_item];
  vector[N_subj] shift_u;  // subj shift
  real<lower=0> shift;
  alpha_c[1] <- add_c + fmax(max(alpha_c_wrong),0);
  alpha_c[2:] <- alpha_c_wrong;
  {
    matrix[N_tau_u,N_tau_u] Lambda_u;
    matrix[2,2] Lambda_rt_u;
    matrix[N_subj,N_tau_u+2]  u_wide;
    Lambda_u <- diag_pre_multiply(tau_u[1:N_tau_u],L_u); #removing the ones
       that go for RTs
    Lambda_rt_u <- diag_pre_multiply(tau_u[(N_tau_u+1):(N_tau_u+2)],L_rt_u);
    u_wide[,1:N_tau_u] <- (Lambda_u * z_u)';
    u_wide[,(N_tau_u+1):(N_tau_u+2)] <- (Lambda_rt_u * z_rt_u)';
    for (i in 1:N_subj){
          for (j in 1:(N_choices-1)){
               u[i,j,] <- u_wide[i,(j-1)*N_coef_u+1:j*N_coef_u];
           }

    u_rt[i] <- u_wide[i,N_tau_u+1];
    u_rt_redo[i] <- u_wide[i,N_tau_u+2];

    }
  }
  {
    matrix[N_tau_w,N_tau_w] Lambda_w;
    matrix[2,2] Lambda_rt_w;
    matrix[N_item,N_tau_w+2]  w_wide;

    Lambda_w <- diag_pre_multiply(tau_w[1:N_tau_w],L_w);
    Lambda_rt_w <- diag_pre_multiply(tau_w[(N_tau_w+1):(N_tau_w+2)],L_rt_w);
    w_wide[,1:N_tau_w] <- (Lambda_w * z_w)';
    w_wide[,(N_tau_w+1):(N_tau_w+2)] <- (Lambda_rt_w * z_rt_w)';
    for (i in 1:N_item){
          for (j in 1:(N_choices-1)){
               w[i,j,] <- w_wide[i,(j-1)*N_coef_w+1:j*N_coef_w];
           }

    w_rt[i] <- w_wide[i,N_tau_w+1];
    w_rt_redo[i] <- w_wide[i,N_tau_w+2];

    }
  }
  alpha_rt <- alpha_rt_raw * logmeanrt;
  shift_u <- shift_u_raw * tau_shift; // =shift_u ~normal(0,tau_shift)
  shift <- shift_raw * logmeanrt;
}

```

```
model {
  P_redo ~ beta(1,1);
  add_c ~ normal(2,2);
  alpha_c_wrong ~ normal(0,2);
  sigma ~ normal(0,1);
  alpha_rt_raw ~ normal(0,1);
  mu_add ~ normal(0,1);
  tau_u ~ normal(0,1);
  tau_w ~ normal(0,1);
  to_vector(z_u) ~ normal(0,1);
  to_vector(z_w) ~ normal(0,1);
  to_vector(z_rt_u) ~ normal(0,1);
  to_vector(z_rt_w) ~ normal(0,1);
  to_vector(beta) ~ normal(0,1);
  tau_shift  ~ normal(0,.5);
  shift_u_raw ~ normal(0,1);
  shift_raw ~ normal(0,1);
  L_u ~ lkj_corr_cholesky(2.0);
  L_w ~ lkj_corr_cholesky(2.0);
  L_rt_u ~ lkj_corr_cholesky(2.0);
  L_rt_w ~ lkj_corr_cholesky(2.0);

  for(i in 1:N_obs){
    vector[N_choices] mu_c;
    real mu_rt;
    real mu_rt_redo;
    real shifted_rt;

    shifted_rt <- rt[i] - exp(shift + shift_u[subj[i]]);
    mu_c[1:N_choices-1] <- alpha_c +  beta * x_betas[,i] +
     u[subj[i]] * x_ut[,i] + w[item[i]] * x_wt[,i] ;
    mu_c[N_choices] <- 0;
    mu_rt <- alpha_rt + u_rt[subj[i]]+ w_rt[item[i]];
    mu_rt_redo <- mu_rt +mu_add + u_rt_redo[subj[i]]+ w_rt_redo[item[i]];

    winner[i] ~ da(shifted_rt,P_redo,mu_c,mu_rt,mu_rt_redo,sigma);
    }
}
```