

# Phantom readings: the case of modified numerals

Paul Marty, MIT  
Emmanuel Chemla, LSCP - ENS - CNRS  
Benjamin Spector, IJN - ENS - CNRS

June 2, 2014

## Abstract

We investigate the mechanisms proposed in formal semantics to account for the ambiguity generated by simple numerical expressions (e.g., ‘three students’). We explain how these mechanisms, when applied to more complex numerical expressions such as ‘between  $n$  and  $m$ ’ (e.g., ‘between three and five students’), predict a surprising ambiguity between a doubly-bounded (e.g., ‘at least three and at most five students’) and a lower-bounded reading (e.g., ‘at least three students’). While the lower-bounded reading is not detectable intuitively, results from three offline experiments and a response time study provide evidence in favor of its existence. Our contribution is twofold. On the experimental side, we present two psycholinguistic methods powerful enough to detect what we call phantom readings, i.e. readings that do not seem to have consequences for actual interpretation, but have detectable effects on processing. On the theoretical side, we show that certain semantic mechanisms that might be thought to overgenerate are in fact vindicated, since they are able to predict certain processing facts that would otherwise remain mysterious. We discuss how these results illustrate the need for a strong integration of formal semantics and psycholinguistic approaches.<sup>1</sup>

## 1 Introduction

A theory of natural language understanding has to include (a) a systematic characterization of the possible meaning(s) of every sentence, based on compositional semantic rules, and (b) hypotheses about how such meanings are computed on-line and accessed by speakers. Formal semantics is traditionally viewed as being concerned with (a), and psycholinguistic research with (b). In this paper, we rely on the case of modified numerals to show the need for a strong integration of these two approaches. For a more general discussion about the relationship between psycholinguistics and linguistics we refer the reader to Lewis and Phillips (2013).

At its core, a formal semantic system for natural languages is an explicit set of rules designed to generate a semantic value for every possible sentence, however complex. A rule introduced for a particular phenomenon often makes predictions that go beyond the original focus of inquiry, which may or may not be empirically adequate. We will show that some semantic rules originally proposed to account for the complex behavior of simple numerical expressions (e.g., ‘three students’), when applied to more complex numerical expressions (e.g., ‘between three and 5 students’), predict what we call *phantom readings*: readings that are invisible to introspection but that can be detected with finer methods.

---

<sup>1</sup>We would like to thank Lewis Bott, Valentine Hacquard, Ira Noveck, François Récanati, Emmanuel Sander as well as audiences at LSCP in June 2011, at MIT (Gibson Lab) in April 2012 and at SALT in May 2012 for useful comments on earlier versions of this paper. We are grateful to Adrian Staub and the anonymous referee for their careful reading of our manuscript and their thorough reviews. Special thanks go to Isabelle Brunet, Anne-Caroline Fiévet and Amanda Swenson for their invaluable practical help. This work has been supported by a ‘Euryi’ grant from the European Science Foundation (‘Presupposition: A Formal Pragmatic Approach’), the European Research Council under the European Union’s Seventh Framework Program (FP7/2007-2013)/ERC Grant Agreement #313610, and by ANR-10-IDEX-0001-02 PSL\* and ANR-10-LABX-0087 IEC. Correspondence concerning this article should be addressed to Paul Marty, Department of Linguistics and Philosophy, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Bldg. 32-D808, Cambridge, MA 02139, USA. E-mail: pmarty@mit.edu

We focus on sentences of the form ‘Between  $n$  and  $m$  As are Bs’, for instance: ‘Between three and five students attended my class’ (hereafter, we will refer to such sentences as *Between* sentences). Intuitively, such a sentence is true just in case the exact number of students who attended my class is 3, 4 or 5. However, as we will see, some influential accounts of the meaning of numerical expressions in natural languages lead us to expect that such a sentence is in fact ambiguous. On top of the reading that we have just described (which we call the ‘doubly-bounded’ reading), this sentence is expected, on theoretical grounds, to have another reading, equivalent to ‘At least three students attended my class’ (we call this reading the ‘lower-bounded reading’). On this reading, the sentence is expected to be considered true if, say, seven students attended my class.

The present paper is organized as follows. First, in the next section, we explain why recent formal semantic views on the meaning of numerals lead to this prediction. Second, we report on three offline experiments which tested subjects’ understanding of *Between* sentences. Their results provide evidence for the existence of the intuitively surprising lower-bounded, ‘at least  $n$ ’-reading. Third, we present data from a response time study. We found a delayed effect in subjects’ response time, which is best understood as an interference of the ‘phantom’, lower-bounded reading of ‘between  $n$  and  $m$ ’ numerals. Taken together, the offline and response time studies strongly suggest that ‘between  $n$  and  $m$ ’ numerals give rise to the ambiguity predicted by formal semantic approaches, even though this ambiguity does not surface at the introspective level. As discussed in the conclusion, these findings show the need for a strong integration of formal approaches, which characterize meaning in representational terms, and psycholinguistic approaches, which are more concerned with the on-line derivation and processing of sentence meaning.

## 2 Existential and maximal readings of numerical NPs

In this section, we explain why a number of formal semantic theories of numerals predict *Between* sentences to be ambiguous between a doubly-bounded reading and a lower-bounded reading. As we will see, the source of this prediction comes from the attempt to account for an attested ambiguity of this sort for simpler expressions. We first describe and motivate two common compositional semantic rules that concern predicates (Existential Closure rule (6)) and bare numerals (Maximization rule (9)). We then discuss how these rules apply to *Between* sentences.

Consider sentences containing a bare numeral, such as (1).

- (1) John has three children.

Out of the blue, (1) means that John has exactly three children. However, in certain contexts, it can be considered true if John has in fact more than three children. Consider for instance the following discourse:

- (2) In order to qualify for a certain tax deduction, one must have three children. *John has three children.* In fact he even has four. So he qualifies for the deduction.

The numeral *three* in (1) thus appears to be ambiguous between a weak, lower-bounded meaning where it is interpreted as meaning *three or more than three*, and a strong, doubly-bounded meaning, where it means *three and no more than three*.

We will now present, in an informal way, a simple account of the ambiguity of bare numerals which is very close in spirit to Geurts (2006), and will show that, with minimal additional assumptions, this account predicts an ambiguity for *between  $n$  and  $m$* . While there is a broad variety of approaches to the ambiguity of bare numerals (see, for a survey, Spector 2013), the tools used by a number of recent influential accounts Breheny (2008); Kennedy (2012), when applied to *modified* numerals such as *between  $n$  and  $m$* , also lead to this prediction. Let us start with the following sentence:

- (3) We are three students.

This sentence is understood to mean that the group of people that the speaker is referring to by *we* (which includes the speaker) has cardinality 3 and is entirely made up of students. Here the numeral *three* is used as a *predicative expression* that represents a property of groups (in the formal semantics literature, such groups are called *plural individuals*), namely the property that a group has if and only

if this group is of cardinality 3. As to the plural noun *students*, it also represents a property of groups, namely the property that a group has if and only if this group is entirely made up of students. The predicative phrase ‘three students’ is then interpreted as representing the *conjunction* of both properties, by virtue of a semantic compositional rule known as *predicate modification*. That is ‘three students’ represents the property that a group has if and only if this group is *both* of cardinality 3 and is made up entirely of students, which seems intuitively correct.

Now consider sentence (4), in which the expression *three students* does not appear as a predicate but in subject position.

- (4) Three students are here.

To get an idea of how to handle this case, it is useful to examine other expressions that can also be used both as predicates and as subject noun phrases. So, let us consider the relationship between the two following sentences:

- (5) a. We are blond students.  
b. Blond students are here.

The sentence (5-a) is standardly analyzed in the same way as (3). The adjective *blond* also represents a property, which can be satisfied by both individuals and groups of individuals (so-called plural individuals). A group of individuals has the property represented by *blond* just in case every member of the group is blond. Now, using the predicate modification rule mentioned above, we know that the phrase *blond students* is going to represent the complex property that a group has if it is both the case that every member of the group is blond and every member of the group is a student. As a result, (5-a) means that the group denoted by *we* has the property represented by *blond students*, namely is a group such that each of its members is a blond student. Turning now to (5-b), we are faced with a difficulty. If we do not say anything more, the subject noun phrase *blond students* represents a property of groups of individuals. But the verb phrase *are here* also represents a property of individuals and groups of individual, namely the property that an individual or a group of individuals has just in case it is located in the vicinity of the author of the sentence. So *are here* normally applies to individuals or groups of individuals, not to *properties* of groups of individuals, hence cannot directly apply to the meaning of *blond students*. This is called a *type mismatch* in formal semantics technical work.

To deal with this problem, one has to assume a special rule of interpretation. We can propose the rule of interpretation in (6), which we call *Existential Closure* (following Geurts (2006) with roots in Partee (1987)):

- (6) Existential Closure  
If *NP* is a noun phrase with no determiner and *VP* is a verb phrase, the sentence *NP VP* is interpreted as meaning that there is an individual or a plural individual *X* such that *X* has both the property represented by *NP* and the property represented by *VP*.

As a result, (5-b) is interpreted as described in (7):

- (7) There exists a group *X* such that *X* has both the property represented by *blond students* and the one represented by *are here*.

This clearly corresponds to the intuitive meaning of (5-b). It says that there are blond students who are here. With this in place, we can go back to bare numerals. Applying the Existential Closure rule to (4) yields the following.

- (8) There exists a group *X* such that *X* has the property represented by *three students* and the property represented by *are here*.

Recall that *three students* represents the property ‘being a group made up of exactly three students’. So (4) is predicted to mean that there exists a group made up of three students that are here.

Now, a crucial observation is that this is just equivalent to ‘Three or more than three students are here’, for the following reason: in any situation where *more than three* students are here, one can find a subgroup of these students of cardinality exactly 3 such that this subgroup is here. This is so because the predicate ‘being here’ is *distributive*: whenever a group of people has the property in question, any

subgroup of this group does too. If we use instead a collective predicate such as *form a circle around the castle*, Existential Closure does not result in a meaning paraphrasable with ‘n or more than n’. Consider for instance *Fifty soldiers formed a circle around the castle*. Existential closure results in ‘There is a group of 50 soldiers and that group formed a circle around the castle’. This proposition is *not* made true by the simple fact that a circle was formed around the castle by 60 soldiers, since in general a proper subpart of a circle is not itself a circle – so that *form a circle around the castle* is not a distributive predicate (see Spector (2013) and the references cited therein for a more detailed discussion).

So what is going on in the case of a distributive predicate is that even though the basic meaning of *three* has an ‘exactly’ component, this component is made invisible by existential closure. The general observation that existential closure can destroy upper-bounded readings in distributive environments is due to van Benthem (1986) and is sometimes called *van Benthem’s problem*. To conclude, the Existential Closure interpretation rule, in the case of bare numerals, leads to the lower-bounded reading when a numerical NP is used as the subject of a distributive verbal phrase (and, in fact, also in object position or any other so-called argumental position, as in ‘Mary met three students’, once the Existential Closure rule is generalized for such cases).

In order to derive the other, doubly-bounded reading, another semantic rule is needed.<sup>2</sup> Following the spirit (but not the letter) of Geurts’ (2006) proposal, we can propose the following rule, which we call the *Maximization Rule*, which specifically targets numerical expressions.<sup>3</sup>

(9) Maximization

A sentence of the form  $[Num\ N]\ VP$ , where *Num* is a numerical expression, *N* is a nominal expression, and *VP* is a verb phrase, is interpreted as meaning that the *maximal* group *X* such that *X* has both the property represented by *N* and the property represented by *VP* has the property represented by *Num*.

Note that, as a nominal expression, *N* may be a bare noun as well as a noun modified by one or several adjectives, a relative clause, or taking a prepositional phrase as a complement. Applied to (4) (‘Three students are here’), this rule gives rise to the following reading: ‘the maximal group that has the property represented by *students* and the property represented by *are here* has the property represented by *three*, i.e. ‘the biggest group consisting of students who are here has cardinality 3’. This, of course, is the desired ‘exact’, doubly-bounded reading. So if our grammar contains both the Existential Closure rule and the Maximization rule, it generates two readings for sentences such as (4), which seems to be a good result, given that bare numerals do give rise to such an ambiguity, as discussed above.

Now, let us see what is predicted for a sentence such as (10).

(10) Between three and five students are here.

Given straightforward assumptions regarding the basic meaning of *between three and five*, it will turn out that applying Maximization as defined in (9) gives rise to the ‘expected’ reading (namely the reading ‘the number of students who are here is comprised between three and five’), but that applying Existential Closure as defined in (6) yields the unexpected lower-bounded reading (‘Three or more than three students are here’).

To see exactly how this comes about, we need first to know what the basic meaning of *between three and five* is when it occurs in a predicative position. Now, it is quite clear that a sentence such as *We are between three and five students* means that the group denoted by *we* is a group of students whose cardinality is between 3 and 5 (inclusively, i.e. is either 3, 4 or 5). So, we want *between 3 and 5* to represent the property that a group has if and only if this group has cardinality 3, 4 or 5. Let us then

<sup>2</sup>For a long time, it was thought that the exact meaning for numerals was derived from the ‘at least’-meaning as a pragmatic, Gricean inference. This view, however, is now believed to be problematic (see, among others, Geurts 2006; Breheny 2008; Spector 2013).

<sup>3</sup>Geurts (2006) proposes a slightly different rule, which amounts to the following:  $[Num\ N]\ VP$  is true iff there is a *unique* group *X* such that *X* has the properties represented by *Num*, *N* and *VP*. When *Num* is a bare numeral and both *N* and *VP* are distributive predicates, this does not change anything. However, for *More than three Ns VPs*, Geurts’ rule gives rise to ‘Four and no more than four *Ns* have the property represented by *VP*’, which is not a desirable result. Replacing *uniqueness* with *maximality* (as in other accounts) solves this problem. From the point of view of this paper, the most important point is that the *Existential Closure* rule, which is assumed in one form or another by all accounts, gives rise to the ‘surprising’ *at least*-reading.

apply Maximization to *Between three and five students are here*, where *between three and five* is the relevant numerical expression (noted as *Num* in the statement of the rule in (9)):

- (11) The maximal group *X* such that *X* both has the property represented by *students* and the one represented by *are here* has the property represented by *between three and five*.

(11) is equivalent to ‘the biggest group consisting of students who are here has cardinality 3, 4 or 5’. This amounts to saying that the number of students who are here is 3, 4 or 5, i.e. is the ‘standard’, expected reading of the sentence.

Finally, let us see what happens if we apply instead existential closure. Here is what we get in this case:

- (12) There exists a group *X* such that *X* has the property represented by *between three and five students* and the property represented by *are here*.

(12) says that there is a group consisting of students whose cardinality is between three and five (inclusively) such that this group is here. Now, similarly to what happened when we applied the rule to (4), we lose the upper-bounded component of ‘between three and five’. This is so because if, say, seven students are here, there exists a subgroup of these students which is made up of exactly four students who are here, which is sufficient to make (12) true. More generally, as soon as there are three or more than three students who are here, one can find a group of students whose cardinality is between 3 and 5 such that this group is here. So by applying the Existential Closure rule to (10), we generate the lower-bounded reading, namely a reading that can be paraphrased by ‘Three or more than three students are here’.

Despite its intriguing predictions, the semantic analysis of *between n and m* developed here has two remarkable advantages. First, it is economic; it relies on the existence of two abstract mechanisms, Maximization and Existential Closure, commonly used as primitives in semantic theories of bare numerals, without any need for extra stipulations. Second, it proposes a uniform account of the meaning of bare and modified numerals. It is worth noting, however, that not all suitable semantic accounts of modified numerals must a priori commit themselves to the predictions we mentioned. There exist other possible theories incompatible with the existence of the phantom, lower-bounded reading we have been arguing for. For instance, it is natural to propose that Maximization is directly built into the lexical meaning of numeral modifiers, leaving no room for Existential Closure to apply. The general idea is that richer numerical expressions such as *between three and five* have, unlike bare numerals, an overt modifier which is assigned a particular piece of meaning, which includes the effects of the Maximization rule. This is in fact the perspective that follows from the standard treatment of comparative determiners such as *fewer than n* (e.g., Heim (2000)). In such an alternative account, Maximization, viewed as optional in the case of bare numerals, should be obligatory in the case of modified numerals.

But, it can be shown that Maximization is not always present with modified numerals of the form *between n and m*. As we noted above in the case of bare numerals, Existential Closure leads to a lower-bounded meaning when the numerical NP is an argument of a distributive predicate but not when it is an argument of a collective predicate. Spector (2014) observes that when a modified numeral of the form *Between n and m* is used in a collective or cumulative context, the reading predicted by Existential Closure is easily detectable. To illustrate, consider the following sentence, which involves the collective predicate ‘form a circle around the castle’ as before or more simply ‘surround the castle’:

- (13) Between thirty and fifty soldiers surrounded the castle.

If Maximization were obligatory, the only possible reading of (13) would be that the maximal group of soldiers that formed a circle around the castle has a cardinality between 30 and 50. Note that, under this reading, (13) is false in a situation where, say, two circles were formed, one with 40 soldiers and the other with 60 soldiers. But (13) seems to have another, quite natural reading, namely that there is group of soldiers with a cardinality between 30 and 50 such that this group formed a circle around the castle. This weaker reading is true in the above situation, and it corresponds to a lower-bounded reading with Existential Closure instead of Maximization. Hence, one can identify sentence-types (e.g., sentences with collective predicates) in which *between n and m* does not involve Maximization. The experimental results reported in this paper further show that even in situations where Maximization seems to be the default option (e.g., distributive contexts), a lower-bounded reading for *between n and m* expressions may still exist. But then we need to explain why this lower-bounded reading is not intuitively available.

Consider again a sentence such as (10), i.e. *Between three and five students are here*. Spector (2014) suggests a pragmatic explanation of the following form: if the speaker intended to convey the lower-bounded reading ('Three or more than three students are here'), she had no need to talk at all about the number 5, which in fact ends up playing no role at all in the meaning of the sentence when it is interpreted by means of the Existential Closure rule. That is, when Existential Closure applies, the identity of  $m$  in *between  $n$  and  $m$*  is irrelevant to the final truth-conditions. When interpreting (10), it is then rational to assume that the speaker did not intend to use the lower-bounded reading. Note that this reasoning does not hold in cases involving collective predicates. For (13), the reading that results from Existential Closure does not make the higher number irrelevant to the sentence truth-conditions. When Existential Closure applies, it does not make (13) equivalent to, say, *Between thirty and **fourty** soldiers surrounded the castle*. If 45 soldiers formed a circle around the castle and no other circle was formed, then (13) is predicted to be true, but not the latter sentence.

To sum up, Existential Closure, when applied to *between  $n$  and  $m$*  in distributive contexts, leads to a reading which, on our view, is ruled out by a pragmatic constraint. However, it remains the case that if the approach we have outlined is correct, the human 'semantic calculator' should generate both the doubly-bounded reading and the intuitively surprising lower-bounded reading for sentences of the form *Between  $n$  and  $m$  VP* (even though the lower-bounded reading is rejected later for pragmatic reasons that go beyond the mere result of semantic computations). In the next sections, we provide behavioral evidence that this is in fact the case, i.e. that naïve subjects do compute this lower-bounded reading.

### 3 Experiments 1, 2 and 3: offline studies

Experiments 1, 2 and 3 were offline studies based on a graded sentence-picture matching task *à la* Chemla and Spector (2011) (see also Marty et al. 2013). Participants were presented with sentence-picture items, as exemplified in Figure 1, and had to assess the extent to which the sentence was a correct description of the situation represented on the picture. Our motivation for collecting graded judgments (as opposed to binary judgments) was to leave room for dispreferred readings to play a role, while binary judgments may be entirely driven by *preferred* readings. For concreteness, we made the following conjecture about graded judgments: for a sentence  $S$  with two distinct readings  $R_1$  and  $R_2$ , participants will judge  $S$  to be true to a higher degree if both readings are true than if only one of them is true, in which case the sentence would still be judged true to a higher degree than if both readings were false.

Experiment 1 was designed to provide a direct test of the predictions made by formal semantic theories regarding the potential readings of *Between* sentences (e.g., 'Between 3 and 5 dots are red'). In the critical cases, these sentences were paired with pictures that make them true under their lower-bounded reading, but false under their doubly-bounded reading. In two control cases, they were paired with pictures that make them either true or false under both readings simultaneously. Following the conjecture outlined above, we expected the ratings to *Between* sentences for the critical cases to be intermediate between the ratings for the two types of controls. To ensure that our interpretation of the task is correct, we carried out two additional experiments. In Experiment 2, we tested a second kind of *Between* sentences, to which current theories only assign doubly-bounded readings (e.g., a French sentence of the form 'The number of dots is comprised between 3 and 5'). In Experiment 3, we tested sentences with bare numerals (e.g., '4 dots are red'), whose ambiguity between a lower-bounded and a doubly-bounded reading has already been discussed in the literature, and may thus serve as a baseline. Our findings provide evidence for the existence of a lower-bounded reading for the first kind of *Between* sentences, and show that this reading is available to a lower degree than for sentences with bare numerals.

#### 3.1 Participants

A total of 36 native speakers of French (25 women), ranging in age from 18 to 33 years (mean age 23 years), participated in these offline studies. For each participant, it was pseudo-randomly determined which one of three experiments they had to complete (Exp.1:  $n=15$ ; Exp.2:  $n=10$ ; Exp.3:  $n=11$ ). The participants gave written informed consent and were compensated for their time. All of them reported to have normal color vision and no prior exposure to formal linguistics.

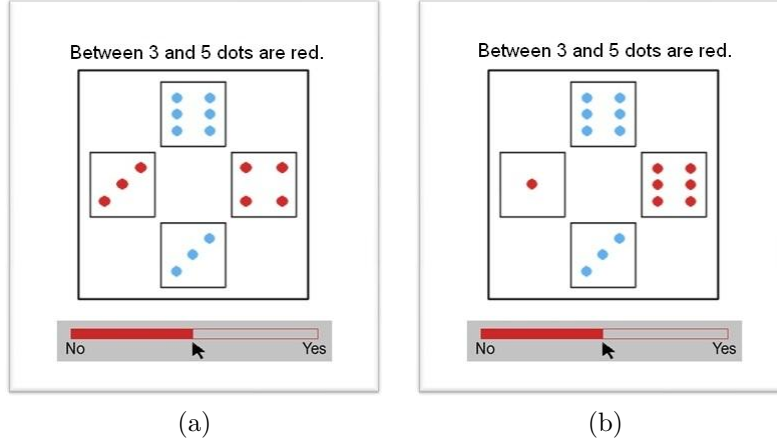


Figure 1: Examples of sentence-picture items. These examples correspond to critical cases for Experiment 1.

### 3.2 Experimental material

Each item consisted of a sentence and a picture (see Figure 1). Sentences used in Exp.1-3 are schematically described in Table 1. Actual sentences were in French, as illustrated below in (14)-(18). Each experiment included one distinct type of target sentences (i.e. *Between*, *Between\** or *Bare Numeral*) and the same two types of control sentences (i.e. *At least* and *At most*). The value of  $n$  varied in the set  $\{3, 4\}$  in Exp.1 and 2, and in the set  $\{3, 4, 5, 6\}$  in Exp.3. The  $\langle \text{color} \rangle$  term was one of the following French color adjectives: ‘rouge’(red), ‘bleu’(blue) or ‘vert’(green).

Exp.	Label	Description of the sentence types
1	<i>Between</i>	Between $n$ and $(n + 2)$ dots are $\langle \text{color} \rangle$ .
	<i>At least</i>	At least $n$ dots are $\langle \text{color} \rangle$ .
	<i>At most</i>	At most $(n + 2)$ dots are $\langle \text{color} \rangle$ .
2	<i>Between*</i>	The number of $\langle \text{color} \rangle$ dots is between $n$ and $(n + 2)$ .
	<i>At least</i>	At least $n$ dots are $\langle \text{color} \rangle$ .
	<i>At most</i>	At most $(n + 2)$ dots are $\langle \text{color} \rangle$ .
3	<i>Bare Numeral</i>	$n$ dots are $\langle \text{color} \rangle$ .
	<i>At least</i>	At least $n$ dots are $\langle \text{color} \rangle$ .
	<i>At most</i>	At most $n$ dots are $\langle \text{color} \rangle$ .

Table 1: Schematic description of the sentence types used in Experiment 1, 2 and 3. For a more concrete illustration, you may read  $n$  as 3 and  $\langle \text{color} \rangle$  as red.

Pictures were composed of four boxes. Each box contained between 1 and 6 dots, which were represented as on the faces of a die to facilitate counting by summing small numbers. In each box, dots were either of the target color used in the sentence (abbreviated to target dots henceforth), or of a different filler color. The  $\langle \text{color} \rangle$  involved in each sentence-picture item was randomly selected from the list of target colors (i.e. red, blue or green). The second color used in the picture was then pseudo-randomly chosen from the list of filler colors (i.e. red, blue, green, purple, yellow, black or gray) minus the selected target color. The number of target dots represented in the pictures varied over the range  $[n - 3, n + 5]$  in Exp.1 and 2, and over the range  $[n - 3, n + 3]$  in Exp.3, giving rise to three types of pictures: *Inferior*, *Intermediate* and *Superior* pictures, as described in Table 2.

In Exp.1, *Between* sentences such as (14) were investigated for their two potential readings:

- (14) Entre 3 et 5 points sont rouges.  
Between 3 and 5 dots are red.

Exp.	Label	Description of the picture
1 & 2	<i>Inferior</i>	$[n - 3, n - 1]$ dots are $\langle \text{color} \rangle$ .
	<i>Intermediate</i>	$[n, n + 2]$ dots are $\langle \text{color} \rangle$ .
	<i>Superior</i>	$[n + 3, n + 5]$ dots are $\langle \text{color} \rangle$ .
3	<i>Inferior</i>	$[n - 3, n - 1]$ dots are $\langle \text{color} \rangle$ .
	<i>Intermediate</i>	$n$ dots are $\langle \text{color} \rangle$ .
	<i>Superior</i>	$[n + 1, n + 3]$ dots are $\langle \text{color} \rangle$ .

Table 2: Schematic description of the picture types used in Experiment 1, 2 and 3, where  $\langle \text{color} \rangle$  and  $n$  refer respectively to the color adjective and the value of  $n$  involved in the sentence they were paired with.

- a. At least 3 and at most 5 dots are red. (doubly-bounded)
- b. At least 3 dots are red. (lower-bounded)

In the critical cases, they were paired with *Superior* pictures that make them true under their lower-bounded reading, but false under their doubly-bounded reading. In the control cases, they were paired with *Intermediate* and *Inferior* pictures that make them respectively true and false under both readings.

In Exp.2, *Between* sentences were replaced with *Between\** sentences, such as (15), in which the modified numeral ‘between  $n$  and  $m$ ’ appears in an environment where it can only yield doubly-bounded readings. We would thus expect participants to judge *Between\** sentences true to a lesser degree than *Between* sentences when paired with *Superior* pictures, since they lack the corresponding lower-bounded reading of *Between* sentences that is true in this critical case.

- (15) Le nombre de points rouges est compris entre 3 et 5.  
The number of red dots is ‘comprised’ between 3 and 5.
- a. Means: The number of red dots is at least 3 and at most 5.
- b. Cannot mean: The number of red dots is at least 3.

In Exp.3, *Between*-sentences were replaced with *Bare Numeral* sentences, such as (16). These sentences complete our range of controls: instead of lacking the potential ambiguity (*Between\** sentences in Exp.2), these sentences uncontroversially generate an ambiguity between a lower-bounded reading (16-a) and a doubly-bounded (16-b) reading.

- (16) 3 points sont rouges.  
3 dots are red.
- a. Exactly 3 dots are red. (doubly-bounded)
- b. At least 3 dots are red. (lower-bounded)

Each experiment also included sentences with the French superlative quantifiers ‘au moins’ (*at least*) or ‘au plus’ (*at most*), as exemplified in (17) and (18) respectively. Lower-bounded and doubly-bounded readings of our target sentences can be paraphrased with these control sentences. Including these *At least* and *At most* sentences on their own thus allows us to evaluate whether participants show any conceptual or practical difficulties in assessing the building blocks of the meaning of interest, when there is no interference of our phenomenon of interest.<sup>4</sup>

- (17) Au moins 3 points sont rouges.  
At least 3 dots are red.
- (18) Au plus 5 points sont rouges.

<sup>4</sup>Note that we are assuming here that *at most* does not itself give rise to ‘phantom readings’ through Existential Closure. However, one could entertain the possibility that the basic meaning of ‘at most  $n$ ’ is ‘being a group consisting in fewer than  $n + 1$  individual’. Then, by applying Existential Closure to (18), we would get ‘there is a group of red dots of cardinality less than 6’, which would be true even when there are 7 red dots. That ‘at most  $n$ ’ is not subject to Existential Closure is clear from the fact that even with collective predicates, the maximality component is present (contrary to what we observed with *between  $n$  and  $m$* ). That is, a sentence such as ‘At most 100 soldiers surrounded the castle’ is false if there is a group of 150 soldiers that surrounded the castle, even if some other group of 50 soldiers also surrounded the castle.



At most 5 dots are red.

Crossing pictures and sentences, we obtained the set of experimental conditions (in bold font) represented in Table 3.

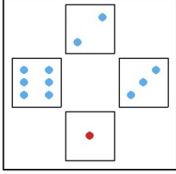
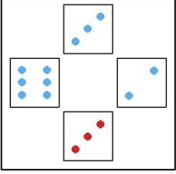
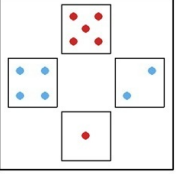
Picture Sentence		<i>Inferior</i>	<i>Intermediate</i>	<i>Superior</i>
				
<b>Target</b>	<i>Between</i> (Exp.1) Between 3 and 5 dots are red.	<b>False</b>	<b>True</b>	<b>Target</b> (true/false?)
	<i>Between*</i> (Exp.2) The number of red dots is between 3 and 5.	<b>False</b>	<b>True</b>	<b>Target</b> (false)
	<i>Bare Numeral</i> (Exp.3) 3 dots are red.	<b>False</b>	<b>True</b>	<b>Target</b> (true/false)
<b>Control</b>	<i>At least</i> At least 3 dots are red.	<b>False</b>	<b>True</b>	<b>Target</b> (true)
	<i>At most</i> At most 5 dots are red.	<b>True</b>	<b>True</b>	<b>Target</b> (false)

Table 3: Summary of the sentence-picture combinations giving rise to the **False**, **True** and **Target** conditions (with  $n = 3$ , and  $\langle \text{color} \rangle = \text{'red'}$  for these examples). Predicted truth-value(s) for the **Target** conditions are given in parenthesis.

### 3.3 Grouping the dots

We worried that participants may restrict their attention to only one of the four dice, as the result of some strategical ‘domain restriction’ or some misunderstanding of the task. Hence, we subdivided the **Target** conditions for the target sentences into two sub-conditions. In the **group** sub-conditions, a *Between* sentence such as ‘Between 3 and 5 dots are red’ was paired with pictures with overall more than 5 red dots, but with (at least) one of the dice containing exactly 3, 4 or 5 red dots. The domain-restriction strategy would thus lead participants to judge the sentence appropriate in this situation, even under their doubly-bounded reading (see example item (a) in Figure 1). Conversely, in the **no-group** sub-conditions, no such groups were available in the pictures (see example item (b) in Figure 1).

Pairwise comparisons between the **group** and the **no-group** sub-conditions were carried out for each type of target sentence. None of these tests reach significance (all  $ts < 1$ ; ns). These results ensure that participants fully understood the task and considered each picture as a whole when judging the sentences, thus ruling out the disrupting strategy we evoked. Hence, the distinction between **group** and **no-group** sub-conditions will be set aside in the remainder of this paper.

### 3.4 Procedure

Sentence-picture items, as illustrated in Figure 1, were displayed in the center of a computer screen using the graphical user interface PyGame for Python. For each item, participants were asked to assess the extent to which the sentence was a correct description of the depicted situation. The participants gave their answers along a continuum of answers, by setting with a cursor the right end of a red line going between ‘No’ (to the left) to ‘Yes’ (to the right). Items remained on the screen until participants entered their answer. The experiment started with 10 trials involving sentences unrelated to the present experimental issue (e.g., ‘There are red dots’): these trials were included to allow participants to familiarize themselves with the paradigm. Next, the participants were presented with test items in a random order,

with a 1500 ms interstimulus interval (blank screen) and two self-timed breaks after each third of the experiment. The number of test items included in Exp.1-3 is given in Table 4. Each experiment was designed so that the proportion of expected ‘true’ and ‘false’ responses was well-balanced. In the **Target** conditions (i.e. *Superior* pictures), one half of the test items for the target *Between*, *Between\** and *Bare Numeral* sentences exemplified the **group** sub-conditions, and the other half exemplified the **no-group** sub-conditions.

Sentence \ Picture	Exp.	<i>Inferior</i>	<i>Intermediate</i>	<i>Superior</i>
<i>Between/Between*</i>	1 & 2	18	36	36
<i>At least</i>		18	36	36
<i>At most</i>		18	36	36
<i>Bare Numeral</i>	3	24	16	48
<i>At least</i>		24	16	48
<i>At least</i>		24	16	48

Table 4: Number of test items included in Experiments 1-3 by experimental condition.

## 3.5 Results

Participants’ responses were coded as the position of the response on the scale, from 0% for a rejection to 100% for acceptance of the sentence as a correct description.

### 3.5.1 Control sentences

Responses to the *At least* and *At most* sentences for the three experiments were as expected: there were small discrepancies for the *At most* sentences that are in line with previous empirical observations,<sup>5</sup> but the overall mean accuracy reached 82% ( $SD = 17$ ). These results show that participants performed the task appropriately.

### 3.5.2 Target sentences

Figure 2 reports the mean ratings to the target sentences by experimental condition (i.e., by picture type). For each of the three experiments, participants’ responses were analyzed using linear mixed-effects regression models (Gaussian family). Each model included Condition as a fixed effect and a maximal random effects structure. In the following, reported  $\chi^2$ -values and  $p$ -values were obtained by performing likelihood ratio tests in which the deviance of a model containing the fixed effect (or interaction) of interest was compared to another model without it, but otherwise identical in random effects structure Barr et al. (2013).

In Exp.1, the mean rating for the *Between* sentences in the **Target** condition ( $M = 34\%$ ) fell between the rating obtained in the **False** condition ( $M = 5\%$ , comparison with **Target**:  $\beta = -29.7$ ,  $t = -3.5$ ,  $\chi^2 = 9.3$ ,  $p < .005$ ) and in the **True** condition ( $M = 84\%$ , comparison with **Target**:  $\beta = 49.4$ ,  $t = 5$ ,  $\chi^2 = 15.4$ ,  $p < .0001$ ). Thus, *Between* sentences were assigned a rating intermediate between ‘clearly false’ and ‘clearly true’ when its two potential readings corresponded to different truth values.

By contrast, in Exp.2, the mean rating for the unambiguous, doubly-bounded *Between\** sentences show a different pattern: the **Target** and the **False** conditions in which the sentence is predicted to be false were not judged differently ( $M = 3\%$  vs.  $M = 2\%$ ,  $\beta = -0.7$ ,  $t = -0.4$ ,  $ns$ ), although both were

<sup>5</sup>Specifically, the global mean score for the *At most* sentences in the conditions where they were expected to be true (i.e. when paired with *Inferior* and *Intermediate* pictures) was notably lower than the score obtained for the *At least* sentences in the relevant corresponding conditions (i.e. when paired with *Intermediate* and *Superior* pictures): 54% vs. 91%. Plausible explanations exist for this discrepancy. Downward-entailing quantifiers like ‘at most’ have been found to be harder to process than upward-entailing quantifiers like ‘at least’ Geurts and van der Slik (2005); Geurts et al. (2010); Cummins and Katsos (2010). The relatively poor performance on the former in our task is consistent with these previous findings and may reflect the same effect.

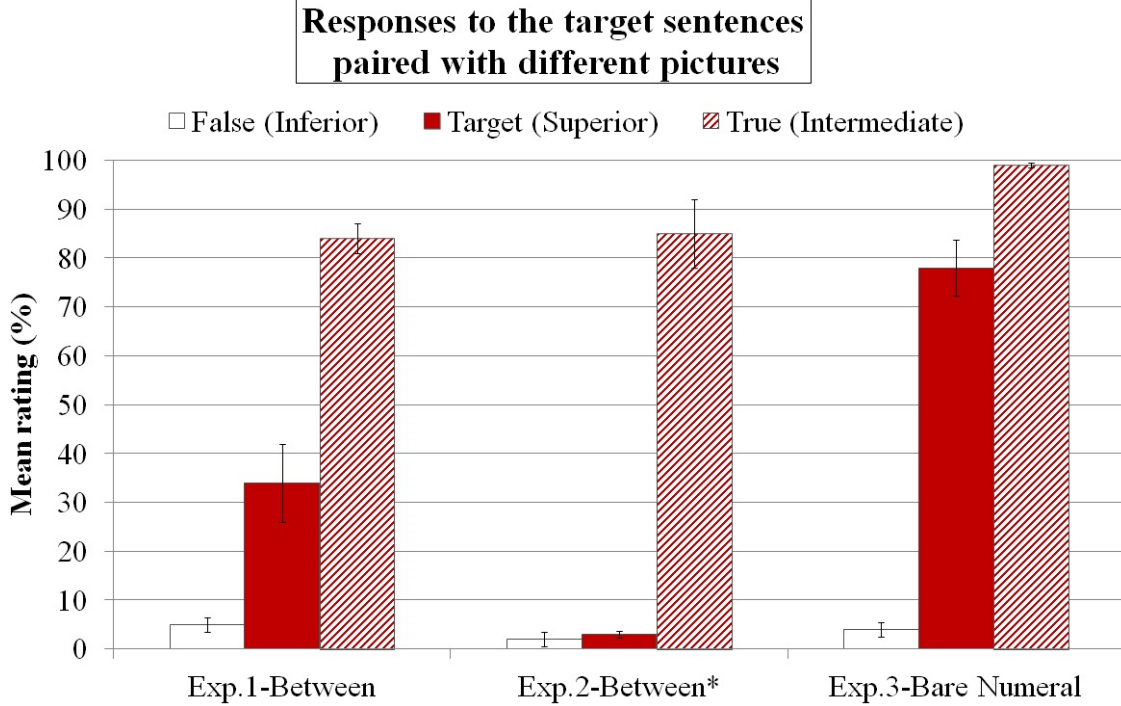


Figure 2: Mean rating (in %) for the *Between*, *Between\** and *Bare Numeral* sentences as a function of the condition (i.e. by picture type). The **Target** conditions (*Superior* pictures) yield intermediate ratings in Exp.1 and Exp.3, i.e. in the two specific cases in which an ambiguity between a true and a false reading is predicted. Error bars refer to standard errors.

judged much lower than the **True** condition ( $M = 85\%$ , all  $\beta$ s  $> 82$ ,  $t$ s  $> 11$ ,  $\chi^2 > 27$ ,  $p$ s  $< .0001$ ). A linear mixed-effects model was fitted in a likelihood setting to examine the effects of Sentence in Exps. 1 vs. 2 and Condition (**False** vs. **Target**) on responses. The model included a maximal random effect structure. There were an effect of Sentence ( $\beta = -32.7$ ,  $t = -3.9$ ,  $\chi^2 = 12$ ,  $p < .005$ ), an effect of Condition ( $\beta = -31$ ,  $t = -3.6$ ,  $\chi^2 = 9.7$ ,  $p < .01$ ), and a significant interaction between Sentence and Condition ( $\beta = 30.2$ ,  $t = 3.5$ ,  $\chi^2 = 9$ ,  $p < .005$ ).

In Exp.3, the ambiguous *Bare Numeral* sentences delivered a three-way pattern similar to the one obtained for *Between* sentences: the mean rating for the **Target** condition ( $M = 78\%$ ) fell between the mean ratings for the **False** condition ( $M = 4\%$ ,  $\beta = -74$ ,  $t = -12.9$ ,  $\chi^2 = 31.5$ ,  $p < .0001$ ) and **True** condition ( $M = 99\%$ ,  $\beta = 21$ ,  $t = 3.5$ ,  $\chi^2 = 8.6$ ,  $p < .005$ ). As above, a linear mixed-effects model was fitted in a likelihood setting to examine the effect of Sentence in Exps. 1 vs. 3 and Condition (**False** vs. **Target**) on responses. The model included a maximal random effect structure. There were an effect of Sentence ( $\beta = 42.2$ ,  $t = 4.2$ ,  $\chi^2 = 10.9$ ,  $p < .005$ ), an effect of Condition ( $\beta = -30.9$ ,  $t = -3.7$ ,  $\chi^2 = 40.5$ ,  $p < .0001$ ), and a significant interaction between Sentence and Condition ( $\beta = -43.1$ ,  $t = -4.2$ ,  $\chi^2 = 10.7$ ,  $p < .005$ ).

This pattern of results is fully explained if we assume that (i) *Between* sentences are ambiguous between a doubly-bounded reading (false in the **Target** condition) and a lower-bounded reading (true in the **Target** condition), just like *Bare Numeral* sentences (albeit to a lower degree) and unlike unambiguous *Between\** sentences, and that (ii) the more readings are true, the higher the sentence is rated.

### 3.5.3 A comparison between Control and Target sentences

In Figure 3, we represent the results for the target sentences in the **Target** condition (*Superior* pictures) along with different baselines: the results for the control *At most* and *At least* sentences in this same condition in each experiment. Participants' responses were fitted into a linear mixed-effects regression

model with Sentence as a fixed effect. All models included maximal random effect structures.

The mean rating for *Between* sentences fell between the *At most* ( $M = 5\%$ ,  $\beta = -20.4$ ,  $t = -2.8$ ,  $\chi^2 = 6.6$ ,  $p < .01$ ) and *At least* ( $M = 93\%$ ,  $\beta = 57.7$ ,  $t = 6.6$ ,  $\chi^2 = 21.4$ ,  $p < .0001$ ) control sentences. The same pattern is found for *Bare Numeral* sentences in Exp.3 (all  $ts > 2.3$ ,  $ps < .05$ ), but not for unambiguous *Between\** sentences in Exp.2 (*At most* vs. *Between\**:  $\beta = 2.7$ ,  $t = 1$ ,  $ns$ ).

Comparisons of responses to the *At most* and *Target* sentences were carried out between Exp.1 and 2, and between Exp.1 and 3, using linear mixed-effects models. Both models included Sentence (*Target* vs. *At most*), Experiment (1 vs. 2 and 1 vs. 3, respectively) and their interaction as fixed effects, and a maximal random effects structure. Results from both models showed an effect of Sentence (all  $\chi^2 > 8$ ,  $ps < .05$ ), an effect of Experiment (all  $\chi^2 > 11$ ,  $ps < .005$ ) and a significant interaction between Sentence and Experiment (all  $\chi^2 > 8$ ,  $ps < .005$ ).

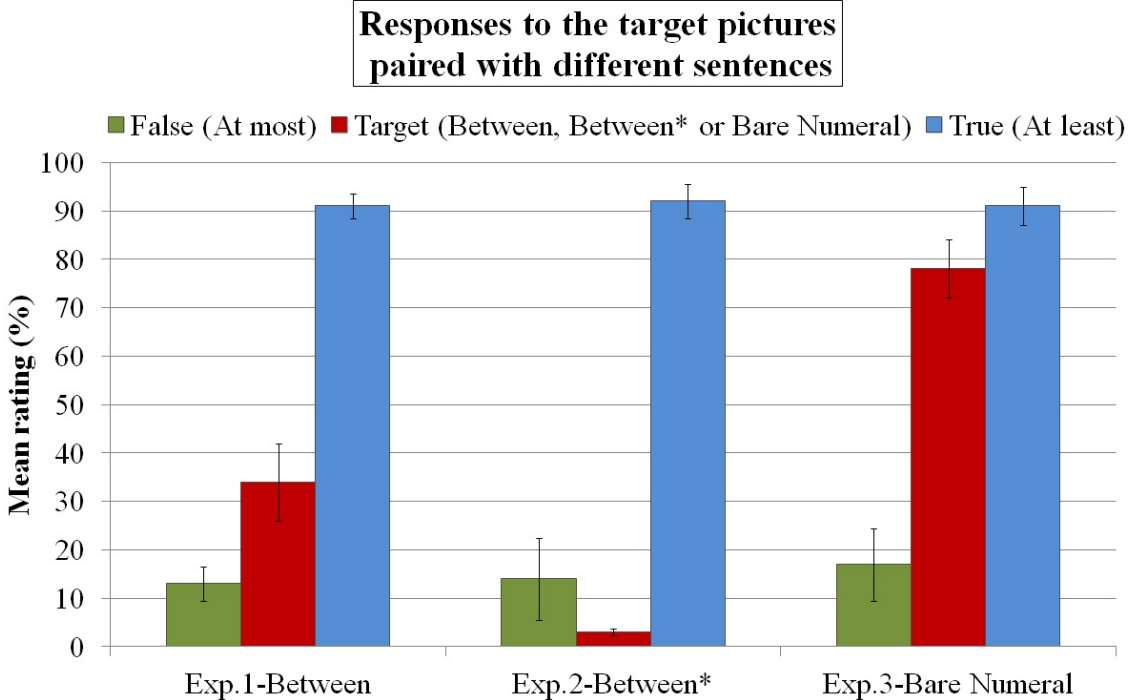


Figure 3: Mean rating (in %) for the *At most*, *Target* and *At least* sentences in the **Target** conditions (*Superior* pictures) as a function of the Experiment. *Target* sentences correspond respectively to *Between* sentences in Exp.1, *Between\** in Exp.2 and *Bare Numeral* in Exp.3 (this part of the results was already presented in Figure 1, along with different baselines). The target sentences yield intermediate ratings in Exp.1 and Exp.3, i.e. in the two specific cases in which an ambiguity between a true and a false reading is predicted. Error bars refer to standard errors.

### 3.6 Discussion of the offline results

We designed a sentence-picture matching task using graded judgments to investigate the meaning of the modified numeral ‘between  $n$  and  $m$ ’. Our results yield three arguments that reveal the existence of what we called a *phantom*, lower-bounded meaning for this expression. The mean rating of *Between* sentences as a description of situations in which the preferred, doubly-bounded reading is false, but the phantom, lower-bounded reading is true has the following properties

**(Result 1)** it is higher than in situations where both the lower-bounded and the doubly-bounded readings are false;

- (**Result 2**) it is higher than similar ratings for *Between\** sentences, which provide otherwise unambiguous glosses of their doubly-bounded reading (but not of their lower-bounded reading);
- (**Result 3**) it is intermediate between ratings obtained for unambiguously true/false *at least* or *at most* control sentences;

We would thus like to interpret the intermediate rating of *Between* sentences in the **Target** condition as support for the claim that this sentence is ambiguous between two readings, one of them being true and the other one being false in that particular condition. However, one may wonder if this intermediate rating could not be explained otherwise. One might suggest that these items are harder to evaluate than others. But it is unlikely to be so for the following reasons. If the *Between* sentences were not ambiguous, the *Between\** sentences would make for an exact paraphrase. But the *Between\** sentences do not behave like the *Between* sentences (Result 2), which we claim shows that (a) they are not ambiguous and (b) it is not particularly difficult to evaluate that the doubly-bounded reading of the *Between* sentences is plainly false in this condition (more precisely, this difficulty is not sufficient to explain the amount of true ratings). One might then like to investigate refinements of the difficulty thesis, which has to involve the particular form of the sentences. One difference between *Between* and *Between\** sentences is that the quantifier is in subject position in the former, and difficulties about quantifiers arise more clearly in subject positions as opposed to object positions (although there is no evidence of this). But the comparison with *At most* sentences is now informative (Result 3): *At most* sentences do not create as much trouble in the **Target** condition, even though the relevant quantifier is now in subject position and has a downward-entailing component (which so far seems to be the source of difficulties for quantified sentences, see Geurts and van der Slik (2005); Geurts et al. (2010); Cummins and Katsos (2010)).

Finally, the fact that the properties above reveal the ambiguity of *Between* sentences was also confirmed by the fact that uncontroversially ambiguous *Bare Numeral* sentences have the same three properties, albeit possibly to a stronger degree. One question we have not yet addressed is why the lower-bounded reading is a phantom reading for *Between* sentences, while it is a visible reading for *Bare Numeral* sentences. Before addressing this question, we present more evidence for the existence of this reading: a delay in the evaluation process of these sentences.

## 4 Experiment 4: time course study

Experiment 4 was based on a two-step sentence-picture matching task, as depicted in Figure 4. Participants were presented with a sentence, e.g., ‘Between 3 and 5 dots are red’, followed by a picture, and had to decide whether the sentence was true or false in the depicted situation. We made the following hypothesis: given two distinct readings  $R_1$  and  $R_2$  for a sentence  $S$ , in cases where  $S$  is true under  $R_1$  but not under  $R_2$ , participants will take more time to answer than in cases where  $S$  is either false or true under both readings.

Note that this hypothesis does not imply full awareness of the ambiguity, but merely requires that different aspects of the stimulus push participants in different directions, making their decision process harder to terminate. Quite generally, assume that stimuli are continually sampled for information, resulting in evidence accumulation over time, until a threshold is reached so that a decision can be issued. If several responses are acceptable, the decision process will receive evidence going both ways, and convergence to a decision will be delayed. This hypothesis is in fact consistent with a wide range of sequential sampling process models of decision making (e.g., Ratcliff et al. 2004; Ratcliff 2002, 1979; Tuerlinckx 2004; Bussemeyer and Rapoport 1988; Vickers 1979; Pike 1966, 1968, 1973) and may be directly linked to classical results in cognitive psychology showing that, when distinct aspects of a given stimulus push subjects in different directions, their decision process takes longer to come out, regardless of the actual outcome. One may think of the Stroop effect as a demonstration of such an interference in response time (cf., Stroop 1935; MacLeod 1991; Roberts and Hall 2008): naming the color in which a color name is written is slower when the color of the text and the color name are incongruent than when they are congruent.

As for the previous offline studies, we constructed target cases in which *Between* sentences were predicted to be true under their lower-bounded reading but false under their doubly-bounded reading, and control cases in which they were either true or false under both of these readings simultaneously. Following the conjecture outlined above, we expected response times to be greater in the target cases than in the other two true/false control cases. To evaluate the assumptions sustaining the interpretation of our

results, we tested two other types of sentences which are well-known to lead to comparable ambiguities, namely sentences with bare numerals (e.g., ‘4 dots are red’) and sentences with scalar items (e.g., ‘Some dots are red’). For each of these three types of sentences, we found that participants’ responses to the target cases were slower than their responses to the control cases.

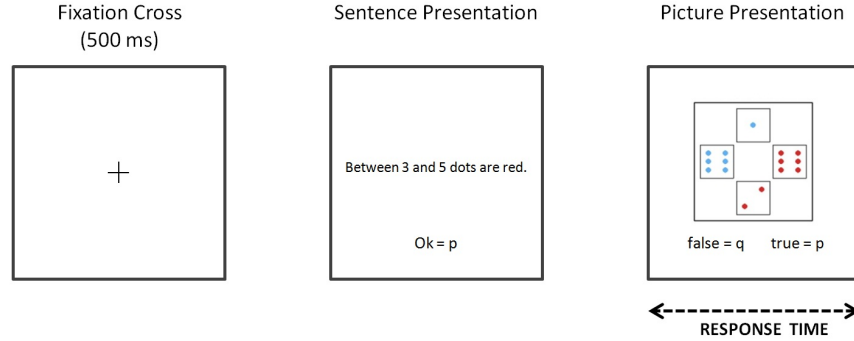


Figure 4: Depiction of the two-step sentence-picture matching task used in Experiment 4.

## 4.1 Participants

33 different native speakers of French (24 women), ranging in age from 19 to 32 years (mean age 23 years), participated in this experiment. The participants gave written informed consent and were compensated for their time. All of them reported to have normal color vision and no prior exposure to formal linguistics.

## 4.2 Experimental material

Each item consisted of a sentence and a picture. Sentences were of the form ‘⟨Quantifier⟩ dots are ⟨color⟩’, where ⟨Quantifier⟩ varied as described in Table 5 (with  $n = 3$  or  $n = 4$ ), and where ⟨color⟩ was a monosyllabic color adjective from a list of target colors (i.e. *red*, *blue* or *green*). Actual sentences were in French, as illustrated below in (19)-(21).

Label	Description of the sentence
<i>Between</i>	Between $n$ and $(n + 2)$ dots are ⟨color⟩.
<i>Bare Numeral</i>	$(n + 1)$ dots are ⟨color⟩.
<i>Some</i>	Some dots are ⟨color⟩.

Table 5: Schematic description of the sentence types used in Experiment 4. For a more concrete illustration, you may read  $n$  as 3 and ⟨color⟩ as *red*.

The general structure of the pictures was the same as in Exp.1-3 (see 3.2 above). The number of target dots represented in the pictures varied as described in Table 6. For the *Between* and *Bare Numeral* sentences, the grouping of target dots was further manipulated into a **group** and a **no-group** sub-condition, as described in 3.3 for the offline studies. The **group/no-group** subdivision will be set aside in the rest of this paper since no difference in participants’ responses and responses times was found between both sub-conditions.

We were interested in comparing the time course of the interpretation of *Between* sentences with those of *Bare Numeral* and *Some* sentences, which provide two genuine phenomena of ambiguity extensively studied in the literature.

- (19) Entre 3 et 5 points sont rouges.  
Between 3 and 5 dots are red.  
a. At least 3 dots are red. (lower-bounded)

Label	Description of the picture	
<i>Inferior</i>	$n - 2$	dots are $\langle \text{color} \rangle$ .
<i>Intermediate</i>	$n + 1$	dots are $\langle \text{color} \rangle$ .
<i>Superior</i>	$n + 4$	dots are $\langle \text{color} \rangle$ .
<i>Null</i>	No	dots are $\langle \text{color} \rangle$ .
<i>Partial</i>	Only some dots are	$\langle \text{color} \rangle$ .
<i>Total</i>	All	dots are $\langle \text{color} \rangle$ .

Table 6: Schematic description of the picture types used in Experiment 4, where  $\langle \text{color} \rangle$  and  $n$  refer respectively to the color adjective and the value of  $n$  involved in the sentence they were paired with.

- b. At least 3 and at most 5 dots are red. (doubly-bounded)
- (20) 4 points sont rouges.  
4 dots are red.
- a. At least 4 dots are red. (lower-bounded)  
b. Exactly 4 dots are red. (doubly-bounded)
- (21) Certains points sont rouges.  
Some dots are red.
- a. At least some dots are red. (lower-bounded)  
b. Some but not all dots are red. (doubly-bounded)

In control cases, sentences were paired with pictures that make both of their readings either true (**True** conditions) or false (**False** conditions). In the target cases, they were paired with pictures that make their lower-bounded reading true, but their doubly-bounded reading false (**Target** conditions). The full list of conditions is given in Table 7. *At least* and *At most* unambiguous sentences from the previous experiments were also included, but only served as filler items for this experiment (they do not give rise to a **Target** condition, necessary for the upcoming analyses).

### 4.3 Procedure

Figure 4 illustrates the procedure. At the beginning of each trial, a sentence was displayed in the center of a computer screen. Participants were asked to press the key ‘p’ as soon as they had read and understood the sentence. The sentence was then replaced by a picture. Participants had to decide whether the sentence was true or false based on the situation represented in the picture, and register their decision by pressing one of two keys (‘q’=false, ‘p’=true). Response times were recorded from picture onset to the point at which the ‘true’ or ‘false’ key was pressed.

Participants were instructed to keep their forefingers on response keys throughout the experiment. They started with a short training composed of 10 trials. Sentences used for the training were unrelated to our experimental purpose (e.g., ‘There are red dots’), and just aimed to familiarize participants with the paradigm. Participants were then presented with 360 sentence-picture items (192 tests and 168 fillers) in random order, with two self-timed breaks. The design was set up so that the proportions of expected ‘true’/‘false’ responses was well-balanced for both test and filler items.

## 4.4 Results

### 4.4.1 Data treatment

In order to minimize the effect of outliers, we trimmed the raw data by removing the items with the 2.5% fastest and 2.5% slowest response times. Response times (RTs) were log-transformed to reduce positive skewness.

Picture Sentence	<i>Inferior</i>	<i>Intermediate</i>	<i>Superior</i>
<i>Between</i> Between 3 and 5 dots are red.	<b>False</b> (12)	<b>True</b> (12)	<b>Target</b> (48)
<i>Bare Numeral</i> 4 dots are red.	<b>False</b> (12)	<b>True</b> (12)	<b>Target</b> (48)
Picture Sentence	<i>Null</i>	<i>Partial</i>	<i>Total</i>
<i>Some</i> Some dots are red.	<b>False</b> (12)	<b>True</b> (12)	<b>Target</b> (24)

Table 7: Summary of the sentence-picture combinations giving rise to the **False**, **True** and **Target** conditions (with  $n = 3$  and  $\langle \text{color} \rangle = \text{'red'}$  for these examples). Numbers in parenthesis refer to the number of test items included in the whole experiment to exemplify the different conditions.

#### 4.4.2 Participants' responses

Figure 5 reports the percent of 'true' responses by experimental condition. Participant's responses to each of the three sentence types were analyzed using generalized linear mixed-effects models for binary data McCullagh and Nelder (1989). Condition was entered in each model as a fixed effect. Random effects for Subject and Item were included, as well as a random slope for Condition grouped by both Subject and Item. For each of the three sentence types, the proportion of 'true' responses in the **False** conditions was lower than in the **Target** conditions (all  $ps < .005$ ), which was lower than in the **True** conditions (all  $ps < .0001$ ). Table 8 provides more information about the output of the statistical models. This first result confirms that all three sentences are ambiguous between a lower-bounded and a doubly bounded meaning.

A generalized linear mixed-effects model predicting responses from Sentence, Condition and their interaction was fitted to the whole response data, and likelihood ratio tests were performed. The full model and all its hierarchical reduced forms included maximal random effects structures. Results showed a main effect of Sentence ( $\chi^2 = 24.7$ ,  $p < .0005$ ), a main effect of Condition ( $\chi^2 = 74.8$ ,  $p < .0001$ ), and a significant interaction between these two factors ( $\chi^2 = 13.3$ ,  $p < .01$ ). This last result shows that the ambiguity is more or less strong for the different phenomena, which was already observed from the results of Exp.1 and 3.

#### 4.4.3 Participants' response times

In the following analysis, we considered RTs for correct responses in the unambiguous **True** and **False** conditions, and RTs for both 'true' and 'false' responses in the ambiguous **Target** conditions. Mean RTs are reported in Figure 6.

The data were analyzed using a linear mixed-effects model predicting RTs from Sentence (*Between*,



Comparison	Sentence	Estimate	Std. Error	z-value	Pr(—z—)
<b>Target</b> vs. <b>False</b>	<i>Between</i>	−1.7	0.5	−3.0	< 0.005
	<i>Bare Numeral</i>	−2.8	0.7	−3.7	< 0.0001
	<i>Some</i>	−9.0	1.5	−6.0	< 0.0001
<b>Target</b> vs. <b>True</b>	<i>Between</i>	6.8	0.6	10	< 0.0001
	<i>Bare Numeral</i>	6.1	0.7	7.8	< 0.0001
	<i>Some</i>	5.3	1.3	4.0	< 0.0001

Table 8: Outputs of the generalized linear mixed-effects regression (glmer) models used to analyze participants’ responses to each sentence type in Experiment 4. Every model included Condition as a fixed effect and a maximal random effects structure.

*Bare Numeral, Some*), Ambiguity (**True/False** vs. **Target**), Response (‘true’ vs. ‘false’) and their interactions. Comparisons of models were then performed based on the Akaike Information Criterion (AIC). The full model and all hierarchical reduced models included maximal random effects structures. The significance of the difference between two models was assessed using chi-squared tests Agresti (2007). A significant increase in residual deviance over the full model was found for Sentence and for Ambiguity: LRT= 28,  $\text{Pr}(\chi^2) < .0001$ , and LRT= 6.8,  $\text{Pr}(\chi^2) < .01$ , respectively. No other factors or interactions was observed to improve the goodness of fit of the model significantly (all LRTs< 3.5,  $\text{Pr}(\chi^2) > .17$ ).

Linear mixed-effects regression models restricted to each sentence type further showed significant effects of Ambiguity: for each sentence type, RTs were significantly longer in the **Target** conditions than in the **True/False** control conditions (all  $ps < .0001$ ). Table 9 provides detailed information about the output of the statistical models.

Sentence	Fixed effect	Estimate	Std. Error	t-value	$\chi^2$	p-value
<i>Between</i>	Ambiguity	$5 \times 10^{-2}$	$6 \times 10^{-3}$	7.8	39.0	< 0.0001
	Response	$1 \times 10^{-2}$	$3 \times 10^{-3}$	4.2	15.4	< 0.0005
	Ambiguity:Response	$-2 \times 10^{-2}$	$1 \times 10^{-2}$	−2.0	3.8	< 0.05
<i>Bare Numeral</i>	Ambiguity	$4 \times 10^{-2}$	$5 \times 10^{-3}$	7.4	36.5	< 0.0001
	Response	$-5 \times 10^{-3}$	$3 \times 10^{-3}$	−1.7	3.1	0.20
	Ambiguity:Response	$9 \times 10^{-3}$	$1 \times 10^{-2}$	0.8	0.6	0.43
<i>Some</i>	Ambiguity	$4 \times 10^{-2}$	$7 \times 10^{-3}$	5.4	21.0	< .0001
	Response	$1 \times 10^{-2}$	$3 \times 10^{-3}$	4.5	17.1	< .0005
	Ambiguity:Response	$-2 \times 10^{-2}$	$1 \times 10^{-2}$	−1.7	2.8	0.09

Table 9: Outputs of the linear mixed-effects regression (lmer) models used to analyze participants’ response time to each sentence type in Experiment 4. Every model included Ambiguity, Response and their interaction as fixed effects, and a maximal random effects structure.  $\chi^2$ -values and p-values were obtained through model comparisons.

In sum, responses to the ambiguous **Target** conditions were slower than responses to the unambiguous **True** and **False** conditions, for each sentence type.<sup>6</sup> These results reveal yet another similarity between the *Between* sentences and the other conditions. They are explained if we assume that (i) *Between* sentences are ambiguous between two readings, each of them leading to a different response in the **Target** conditions, and that (ii) the more responses are acceptable, the longer the response time is. On the other hand, they would not be expected if only the doubly-bounded readings were generated for any

<sup>6</sup>Under the hypothesis that correct responses might be faster than incorrect responses, one may wonder whether the observed effects of Ambiguity could be partly driven by the current data treatment. For, while it is *in principle* possible to distinguish correct responses from errors in the unambiguous **True/False** conditions, no such a distinction can be made in the ambiguous **Target** conditions, where both response types are assumed to be acceptable. Hence, to rule out this alternative explanation, we carried out an extra analysis that considers RTs for all response types in both ambiguous and unambiguous conditions. Results from this second analysis yielded the same conclusions as the ones currently reported.

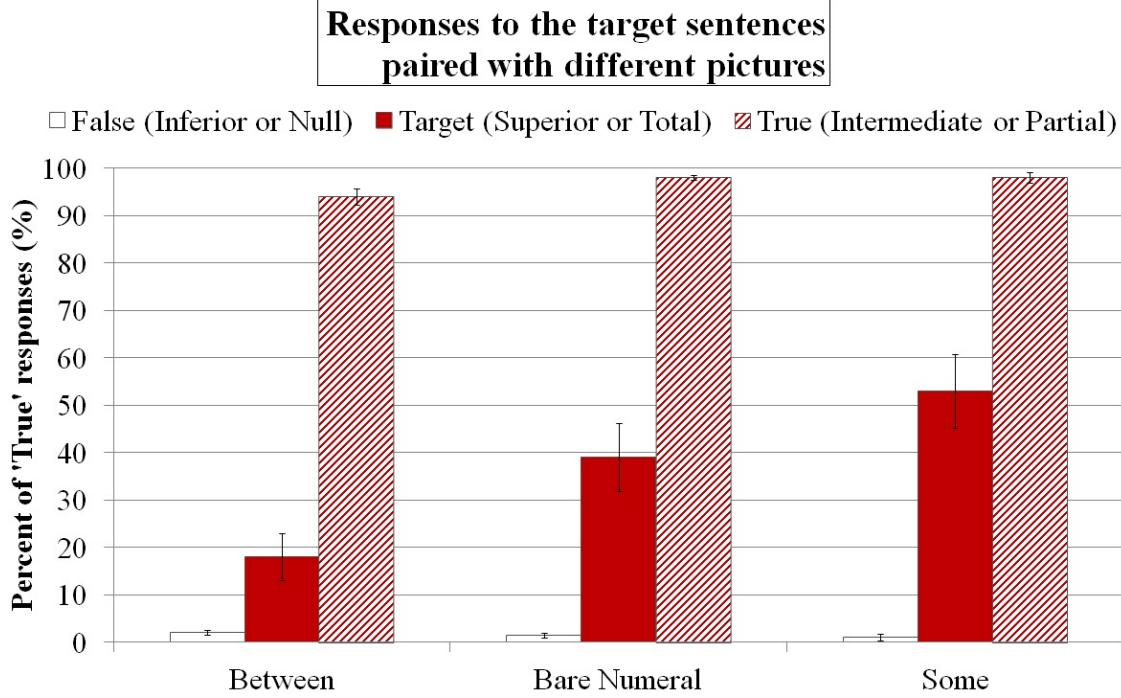


Figure 5: Percent of ‘true’ responses (%) for the *Between*, *Bare* and *Some* sentences as a function of the condition (i.e. **False**, **Target** and **True**). Error bars indicate 95% confidence intervals estimated from binomial distributions.

of these sentences.

#### 4.5 Discussion of the online results

We designed a two-step sentence-picture matching task to compare the duration of subjects’ decision process to sentences quantified with ‘between  $n$  and  $m$ ’, bare numerals and ‘some’. We reasoned that if the modified numeral ‘between  $n$  and  $m$ ’ leads to ambiguous readings, comparable to those generated by bare numerals and scalar items like ‘some’, we should observe a similar RT pattern. Specifically, we obtained greater response times in the target conditions, in which the two readings push towards different responses, which is explained if the two readings interfere with each other. These findings indicate that participants’ decision process was harder to terminate in cases where the sentence-picture combination is predicted to yield conflicting responses.

Arguably, pictures may be more or less hard to scan depending on the number of target dots represented on them. Nonetheless, the present results cannot be accounted for only in terms of the properties of the pictures, that is independently of the sentence they were paired with. Indeed, if the RT differences found between the target and the control conditions were simply due to the fact that participants had more dots to count on pictures of the former than of the latter, similar RT differences should have been observed between both control conditions (since participants also had more dots to count in the true than in the false control conditions). Furthermore, no such difference in RTs should have been found for the scalar *some*-sentences whose verification does not require participants to count the dots in order to provide a truth-value judgment. Hence, these processing results confirm the findings from Experiments 1 to 3, and provide new empirical support for the ambiguity of ‘between  $n$  and  $m$ ’.

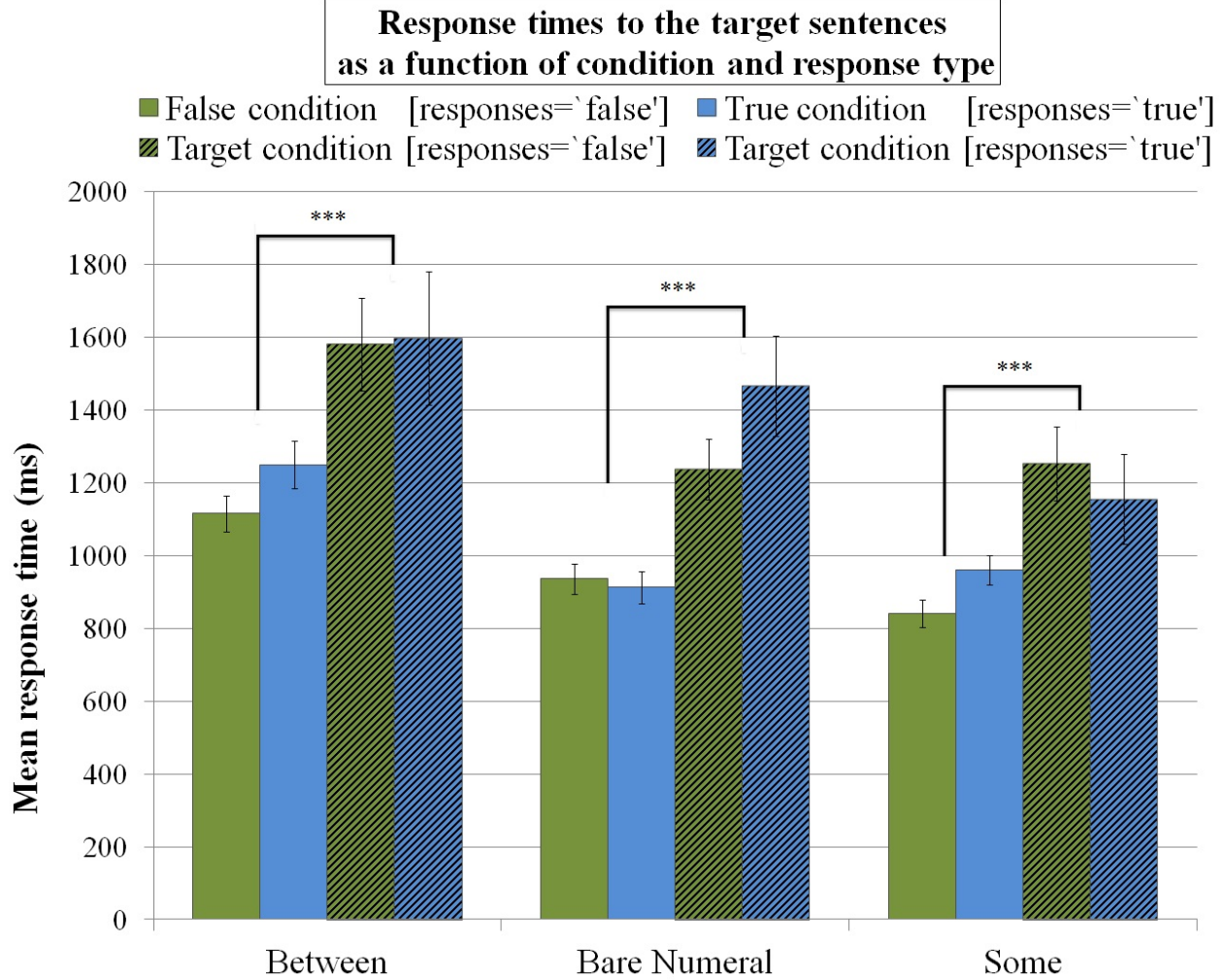


Figure 6: Mean RT (in ms) to the *Between*, *Bare Numeral* and *Some* sentences for correct responses in the **True** and **False** conditions, and for both ‘true’ and ‘false’ responses in the **Target** conditions. The first two bars are always shorter than the last two. Error bars refer to standard errors.

## 5 Conclusion

Formal theories posit semantic mechanisms which are supposed to apply in full generality, that is beyond the basic set of data that they are originally designed to account for. It is the possibility of extending those mechanisms to novel sentences that makes these theories appealing and fully testable. In this paper, we explained how the mechanisms hypothesized to account for the ambiguity of bare numerals between a lower-bounded reading and a doubly-bounded reading are expected to yield a similar kind of ambiguity for more complex numerical expressions of the form ‘between  $n$  and  $m$ ’. While these expectations seemed to be at odds with introspective judgments, we provided evidence that the initial theories were correct and can account for behavior beyond mere introspection. In short, we found that if a phantom reading has a different truth-value than the salient reading, then (a) it affects a particular kind of truth value judgment, and (b) it generates a measurable delay in response time.

Our data also reveal that, beyond similarities in their qualitative patterns, the interpretive processes involved in resolving the ambiguities at hand respond to different forces. Even though naïve subjects do have access to the lower-bounded reading of *between  $n$  and  $m$*  expressions, our results indicate that

it is not salient. More specifically, the lower-bounded reading is less salient for *Between* sentences than it is for bare numerals, despite the fact that these readings are generated by the same means. We think however that it is these differences that should receive new explanations, along the following lines. Assuming that ‘Between 3 and 5 people came’ is ambiguous, as we do, there are reasons why this sentence will nonetheless *not* be used to convey its ‘at least’ reading. First, there are much simpler expressions that could convey this ‘at least’ reading, e.g., ‘At least 3 people came’, which also contains a modified numeral (albeit a simpler one, if we follow, e.g., the complexity measure from Katzir (2007)) or even a bare numeral as in ‘3 people came’, which can also receive an ‘at least’ interpretation. Another oddity of this ‘at least’ reading for *Between* sentences, which is another version of the same argument though, is that the numeral ‘5’ seems to be completely idle: it could be replaced with any numeral (above 3). The resulting ‘at least’ meaning is unaffected by this part of the sentence. Arguably, this discussion is not yet a formalized model of which reading may remain unused in everyday life, that is which readings are *phantom readings*. However, our hope is that we have provided convincing arguments towards why the specific reading we were after is a phantom reading. We hope that a more general theory of phantom readings can be developed, and help us think about apparent cases of misalignment between semantic models and introspective judgments from a new perspective.

At the interface between linguistics and psycholinguistics, our work argues in favor of a stronger integration of formal semantics approaches, which characterize meaning in representational terms, and experimental approaches, which are more concerned with the on-line derivation and processing of sentence meaning. On the one hand, psycholinguistic methods are necessary to fully evaluate the fine-grained predictions made by formal linguistic theories and for which there is no direct introspective evidence. On the other hand, the relevance of experimental results can only be assessed relative to formally explicit models of sentence meaning. The present results illustrate that theoretical linguistic inquiries are required to provide a proper interpretation of processing facts.

## References

- Agresti, A. (2007). *An introduction to categorical data analysis*, Volume 423. Wiley-Interscience.
- Barr, D. J., R. Levy, C. Scheepers, and H. J. Tily (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language* 68(3), 255–278.
- van Benthem, J. (1986). *Essays in logical semantics*, Volume 29 of *Studies in Linguistics and Philosophy*. Reidel.
- Breheny, R. (2008). A new look at the semantics and pragmatics of numerically quantified noun phrases. *Journal of Semantics* 25(2), 93.
- Bussemeyer, J. R. and A. Rapoport (1988). Psychological models of deferred decision making. *Journal of Mathematical Psychology* 32(2), 91–134.
- Chemla, E. and B. Spector (2011). Experimental evidence for embedded scalar implicatures. *Journal of Semantics* 28(3), 359–400.
- Cummins, C. and N. Katsos (2010). Comparative and superlative quantifiers: Pragmatic effects of comparison type. *Journal of Semantics* 27(3), 271.
- Geurts, B. (2006). Take ‘five’: the meaning and use of a number word. In S. Vogeeler and L. Tasmowski (Eds.), *Non-definiteness and Plurality*, pp. 311–329. Amsterdam/Philadelphia: John Benjamins.
- Geurts, B., N. Katsos, C. Cummins, J. Moons, and L. Noordman (2010). Scalar quantifiers: Logic, acquisition, and processing. *Language and Cognitive Processes* 25(1), 130–148.
- Geurts, B. and F. van der Slik (2005). Monotonicity and processing load. *Journal of semantics* 22(1), 97.
- Heim, I. (2000). Degree operators and scope. In B. Jackson and T. Matthews (Eds.), *Proceedings of SALT*, Volume 10, Ithaca, NY, pp. 40–64. Cornell University.

- Katzir, R. (2007). Structurally-defined alternatives. *Linguistics and Philosophy* 30(6), 669–690.
- Kennedy, C. (2012). A scalar semantics for scalar readings of number words. Ms., University of Chicago.
- Lewis, S. and C. Phillips (2013). Aligning grammatical theories and language processing models. *Journal of Psycholinguistic Research*. in press.
- MacLeod, C. M. (1991). Half a century of research on the stroop effect: an integrative review. *Psychological bulletin* 109(2), 163.
- Marty, P., E. Chemla, and B. Spector (2013). Interpreting numerals and scalar items under memory load. *Lingua* 133, 152–163.
- McCullagh, P. and J. A. Nelder (1989). Generalized linear models (monographs on statistics and applied probability 37). *Chapman Hall, London*.
- Partee, B. (1987). Noun phrase interpretation and type-shifting principles. In J. Groenendijk, D. de Jong, and M. Stokhof (Eds.), *Studies in discourse representation theory and the theory of generalized quantifiers*, pp. 115–143. Dordrecht: Foris.
- Pike, R. (1966). Stochastic models of choice behaviour: Response probabilities and latencies of finite markov chain systems<sup>1</sup>. *British Journal of Mathematical and Statistical Psychology* 19(1), 15–32.
- Pike, R. (1968). Latency and relative frequency of response in psychophysical discrimination. *British Journal of Mathematical and Statistical Psychology* 21(2), 161–182.
- Pike, R. (1973). Response latency models for signal detection. *Psychological Review* 80(1), 53.
- Ratcliff, R. (1979). Group reaction time distributions and an analysis of distribution statistics. *Psychological bulletin* 86(3), 446.
- Ratcliff, R. (2002). A diffusion model account of response time and accuracy in a brightness discrimination task: Fitting real data and failing to fit fake but plausible data. *Psychonomic Bulletin & Review* 9(2), 278–291.
- Ratcliff, R., P. Gomez, and G. McKoon (2004). A diffusion model account of the lexical decision task. *Psychological review* 111(1), 159.
- Roberts, K. L. and D. A. Hall (2008). Examining a supramodal network for conflict processing: a systematic review and novel functional magnetic resonance imaging data for related visual and auditory stroop tasks. *Journal of cognitive neuroscience* 20(6), 1063–1078.
- Spector, B. (2013). Bare numerals and scalar implicatures. *Language and Linguistics Compass* 7(5), 273–294.
- Spector, B. (2014). Plural Indefinites and Maximality. *Talk at the UCLA linguistics colloquium*.
- Stroop, J. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology* 18(6), 643.
- Tuerlinckx, F. (2004). The efficient computation of the cumulative distribution and probability density functions in the diffusion model. *Behavior Research Methods, Instruments, & Computers* 36(4), 702–716.
- Vickers, D. (1979). *Decision processes in visual perception*. Academic Press New York.