Abstract

What is the semantic content of a question? As pointed out by Karttunen (1977), declarative sentences that embed interrogative complements (such as “John knows which students called”) can provide relatively easy access to the semantics of questions. Recent theories attribute different readings to such sentences and their predictions depend in various ways on the embedding verb (‘know’ in this example). Through a series of four experiments, we provide quantitative offline data to evaluate critical judgments from the literature. We show that the so-called strongly exhaustive reading is not the only available reading for ‘know’, providing an argument against approaches inspired by Groenendijk and Stokhof (1982, 1984). We also describe processing data which may further constrain theories, provided hypotheses about the derivation processes are made explicit.

1 Introduction

1.1 The meaning of questions

Our goal is to provide quantitative data to help decide between different semantics which have been proposed for questions. We cannot study the meaning of questions, such as (1), as we traditionally study the meaning of declarative sentences, such as (2). Declarative sentences can be described by their truth conditions (Frege, 1892; Tarski, 1935, 1956). Knowing the meaning of (2) may be reduced to knowing which situations make this sentence true and which situations make it false. For questions, we cannot define truth conditions. It would not make any sense to say that (1) is true or false in a given situation.

(1) Which students called?

---

*We would like to thank Benjamin Spector, Paul Egré, Manuel Kriz, Paul Marty, Salvador Mascarenhas and Lyn Tieu for discussion and important comments, Seth Cable as an editor and two anonymous reviewers for their detailed comments on an earlier version, Nicolas Porot and Aparicio Kozuch for their help with the English materials, Radhia Achheb for her administrative support, Alex Drummond for the IBEX interface and his help using it, and the audiences at the XPrag London Masterclass and the Euro-XPrag 2013 Conference in Utrecht for their attention and comments. The research leading to these results has received funding from the European Research Council under the European Union’s Seventh Framework Programme (FP/2007-2013) / ERC Grant Agreement n.313610 and was supported by ANR-10-IDEX-0001-02 PSL* and ANR-10-LABX-0087 IEC.

† CNRS, LSCP, ENS Paris
Mary called.

Hamblin (1958, 1973) proposed that the meaning of a question resides in its answerhood conditions. Knowing the meaning of (1) is equivalent to knowing what counts as an answer to (1). Because answers are propositions, with truth conditions, questions can be studied within the general framework of truth-conditional semantics. The problem is then reduced to identifying the set of possible answers to a question. This tells us what to look for, but not how to look for it.

1.2 Embedding questions as a way to study them

Some verbs such as know can embed questions, as in (3), just as well as declarative sentences, as in (4).

(3) John knows which students called.
(4) John knows that Mary called.

As Karttunen (1977) pointed out, the meaning of (3) must be related to the meaning of the question (1) (and similarly (4) must be related to (2)). More specifically, (3) seems to be true in situations where John is able to answer the question (1). Because (3) is a declarative sentence, it can receive a truth value and we know how to study it. Thereby, sentences like (3) provide some access to the semantics of the question (1). As we will see, this provides an entry point into understanding questions both from a theoretical and from a psycholinguistic point of view.

Among all the predicates that can embed questions, we will focus on those like know, which are called veridical responsive predicates. As we saw, these verbs can embed both interrogative and declarative complements.¹ They are interesting because the meaning of (3) is usually assumed to be reducible to something of the form “John knows that p” for some proposition p (namely, an answer to the question). If this is the case, everything we know about the declarative-embedding know can be applied to the question-embedding know.²

1.3 Different readings for embedded questions

Let us first review the bare facts. Several readings have been associated with sentences such as (3) in the literature. We can distinguish between readings which are called exhaustive and those which are called non-exhaustive. In the first part of this section we will present the different exhaustive readings, which are the focus of our study. Their distribution has been strongly debated in the theoretical literature, and our main goal will be to provide quantitative data that speak to their respective distributions. The second part of this section presents an example of a non-exhaustive reading. Non-exhaustive readings will not be

---

¹Predicates which only embed interrogative complements, such as wonder, are called rogative. Non-veridical responsive predicates, such as agree, can receive both types of complements. They differ from know in that they can express a relation to any answer to the question, not only to the true answer. For instance, it may be true that John and Peter agree on which students called while both are wrong.

²This reducibility property should not be taken for granted. See Schaffer (2007) and George (2011) for discussion and Chemla and George (2013) for experimental evidence supporting non-reducibility for the non-veridical responsive predicate agree.
given a prominent role here (although we will briefly discuss their potential role in our first experiment).

**Exhaustive readings:** The different exhaustive readings of (3) described in the literature are presented in (5). They are all called *exhaustive* readings because they require that John have complete knowledge about the students who actually called. They differ on the required knowledge about students who did not call. The Strongly Exhaustive (SE) reading described in (5a) requires complete knowledge about both callers and non-callers. It entails the Intermediate Exhaustive (IE) reading (5b), which merely requires absence of false beliefs about non-callers. The IE reading in turn entails the Weakly Exhaustive (WE) reading (5c), which does not require anything beyond ‘exhaustivity’, that is true beliefs about the students who actually called.

(5) Exhaustive readings of (3):
   a. Strongly exhaustive (SE):
      For each student who called, John knows that she called, and he knows that no other student called.
   b. Intermediate exhaustive (IE):
      For each student who called, John knows that she called, and John does not have false beliefs about students who didn’t call.
   c. Weakly exhaustive (WE):
      For each student who called, John knows that she called.

The existence and co-existence of these readings is debated, judgments vary across theoreticians and certainly across verbs. Klinedinst and Rothschild (2011) present the only quantitative survey of the issue, which results suggest that different native speakers may provide responses coherent with any of the three exhaustive readings for questions embedded under the verb *predict.*

**Non-exhaustive readings:** Although they are not our focus, non-exhaustive readings have also been discussed. At least for *know*, they only seem to be available under specific circumstances. As an example, (6), repeated from George (2011), is usually considered true as soon as William is able to name at least one place where Rupert could buy an Italian newspaper. Any exhaustive reading seems too strong, since it would require that William know all the places where Italian newspapers can be purchased. This non-exhaustive reading is usually called the *mention-some* reading. It will not be at the foreground of our inquiry.

(6) William knows where Rupert can buy an Italian newspaper.

---

3A first description of this intermediate reading can be found in Groenendijk and Stokhof (1982), but the authors quickly reject it in favor of the SE reading (thanks to Benjamin Spector for pointing this out). The intermediate reading resurfaces in Spector (2005) (and possibly Berman, 1991). Certainly, the IE reading has been implemented in very few formal approaches.
1.4 Architecture of recent theories of questions and embedded questions

We cannot offer a complete introduction to the theories of questions, but in this section we try to give a sense of what they are made of and how they may vary. The main material at the basis of our experimental inquiries is to be found in the empirical predictions of the relevant theories. We introduce these empirical aspects independently of their formal and conceptual basis in the following section.

In this section, we focus on theories which follow the general architecture originally defined by Hamblin (1973). These theories are based on three essential ingredients: (a) a denotation of questions which corresponds to a set of answers/propositions, (b) some mechanism to combine such sets of propositions with verbs which normally combine with propositions and possibly (c) some form(s) of strengthening mechanism.

(a) Answer sets. A theory of questions first owes a basic denotation for a question. This is typically a set of answers. We can define two groups of theories, depending on which set of answers is taken as the denotation of the question.

- On the first group of theories, following Hamblin (1973) and Karttunen (1977), the denotation of a question consists of so-called ‘weak answers’, which can be defined through an example:

  (7) Weak answers for a question such as *Who called?:*  
  Mary called, Peter called, Mary and Peter called...

  As a result of the combination of this denotation with the compositional mechanism (see (b) below), the WE reading is primitive. The other readings may then be derived by applying strengthening mechanisms (see (c) below). Most recent theories fall into this group (Beck and Rullmann, 1999; Lahiri, 2002; Sharvit, 2002; Guerzoni, 2003, 2007; Guerzoni and Sharvit, 2007; Klinedinst and Rothschild, 2011; Spector and Egré, 2014).\(^5\)

- On the second group of theories, the denotation of a question consists of so-called ‘strong answers’, see (8), as in Groenendijk and Stokhof (1982, 1984, 1993) and more recently George (2011).

\(^4\)We do not discuss theories which strongly divert from the work of Hamblin (1973). Some approaches, such as Ginzburg (1996), put less stringent constraints on the set of possible readings and focus their inquiry more on the role of context in the selection of a reading. The inquisitive semantics framework has also given rise to much work on questions, partly in reaction to the limitations of partition semantics (see Mascarenhas, 2009), but we are not aware of any explicit implementation of question embedding in an inquisitive framework. Nevertheless, the main conclusion from our study applies to these theories as well: they must all be able to derive some form of the intermediate exhaustive reading.

\(^5\)Strictly speaking, Spector and Egré (2014) use a rule which quantifies over strong answers, so they should fall into the second group. However, they also refer to the weak answers in order to derive the IE reading. Because weak answers cannot be retrieved from the strong ones, it makes sense to consider their theory as a member of the first group.
(8) Strong answers for a question such as *Who called?*

Only Mary called, Only Peter called, Only Mary and Peter called...

As a result, the SE reading is primitive. If we allow only strengthening mechanisms, as opposed to weakening mechanisms, the other exhaustive readings are simply not available.6

(b) **Combination with responsive verbs.** This is usually the most technical part of a theory: it formalizes how compositional semantics combines the denotation of a question (set of answers, see (a) above) with the semantics of the verb with which it merges. A simple example would be the following rule for the verb *know*: “knowing \( Q \) is equivalent to knowing the conjunction of all true answers to \( Q \)”. So, if \( Q = \{a_1, a_2, a_3\} \) and \( a_1 \) and \( a_2 \) are true but \( a_3 \) is false, then \( [\text{know } Q] = [\text{know } a_1 \land a_2] \).

Actual theories usually provide a uniform treatment of all responsive verbs by means of a generic abstract rule. These rules are somewhat complex and there is room for variation. For instance Lahiri (2002) considers that the verb is lexically responsible for restricting (or not) the set of relevant answers, e.g., to true answers; Spector and Egré (2014) do not use universal quantification but rather existential quantification (roughly: \( [\text{know } Q] = \exists a \in Q : [\text{know } a] \)), and they also treat the assertive and presuppositional dimensions independently.

(c) **Strengthening mechanisms.** Various semantic and pragmatic mechanisms have been described for strengthening the meaning of an utterance.

- One example is exhaustification, through a grammaticalized EXH-operator as in e.g., Fox (2007) or as a proxy for Gricean strengthening. For our purposes, the potential application of exhaustification can help derive stronger readings from weaker ones, even though weaker readings remain primitives.

- Another example more specific to the theory of questions is given by Heim (1994), who showed that the set of strong answers could be recovered from the set of weak answers (but not *vice versa*). (Note that this role can also be supported by a more general EXH operator as described above, provided that it can be applied at the level of individual answers in the denotation of the question). Such a mechanism effectively makes the first group of theories as described in (a) above capable of encompassing the predictions of the second group of theories.

Let us present a semi-concrete theory that puts together all of these ingredients. We could consider a theory which (a) makes use of weak answers, (b) relies on the ‘universal’ embedding rule and (c) allows for the EXH-operator to produce stronger readings (as a first approximation, this operator strengthens a proposition by conjoining it with the negation of all its non-weaker alternatives). Such a theory would be very close to the approach of

---

6George (2011) makes use of the weak answers to derive mention-some readings (non-exhaustive) but his theory cannot derive WE readings. He argues that alleged cases of WE readings, such as the examples of Berman (1991) with the verb *surprise*, are in fact mention-some readings.
Klinedinst and Rothschild (2011) and would be able to derive all of the exhaustive readings that we discussed. First, WE readings would be primitives (using weak answers and no exhaustification yields a weak reading). Second, IE readings would be obtained by exhaustification at the matrix level (see Klinedinst and Rothschild, 2011 for details). Finally, SE readings would be obtained by local exhaustification, the reason being roughly that this would turn weak answers into strong answers, in the sense of (7) and (8), as suggested below:

\[
\text{exh}('Peter and Mary called', \text{alternatives} = \{\text{‘Sue called’, ‘Jack called’, ...}\}) = \\
[Peter and Mary called] \land \neg[Sue called] \land \neg[Jack called] \land ...
\]

The combinations of various versions of the three ingredients we presented in this section lead to a variety of possible theories. In the next section, we will focus on the empirical predictions of the current implementations found in the literature. In short, starting with weak readings allows for more flexibility, because various tools can be called for to obtain strengthened meanings from weaker ones (e.g., exhaustification, maximization) and less so in the other direction.\(^7\) Although we will not enter into the details of which theories rely on which tools, we will describe the predictions they make.

1.5 Predictions of recent theories for know

Let us put aside the technical differences between the different theories and the way they derive the different readings. Here we merely want to provide the set of readings each theory predicts to be available, so that we can later evaluate whether these predictions are borne out. Let us also focus on the verb know, which has been discussed in greater detail than any other responsive verb. We thus want to summarize which readings are available for know, according to some recent theories from the two groups identified in 1.4(a). This is done in Table 1, which indicates whether a given theory can derive a given reading or not.

Let us note that theories also vary as to whether they offer different predictions for different verbs. The comparison between different verbs is in fact crucial to put constraints on the available compositional mechanisms between the verb and the embedded question (see 1.4b). However, this cross-verb variation will not be the focus of our inquiry, which focusses on know, and we therefore do not report a complete set of predictions across verbs. Let us immediately note however that in Experiment 2 we also tested the verb predict, and that the results did not reveal important differences with know.

1.6 Goals of our study

The goal of our study will be to provide constraints on what readings an accurate theory should derive and how. We will test the availability of the different exhaustive readings and compare these results with the predictions in Table 1. We will also gather online data which

\(^7\) Furthermore, Heim (1994) proved that under some hypotheses (constant domain) it is logically impossible to retrieve the weak answers from the strong ones.

\(^8\) Klinedinst and Rothschild (2011) argue against the availability of WE readings for verbs like know, but their restriction is not derived by the theory; rather it is a further pragmatic constraint that the authors stipulate. As it stands, their theory derives IE readings for non-factive verbs only, but they indicate how it could be modified in order to derive IE readings for all veridical responsive verbs.
Table 1: Theories make different predictions about which (exhaustive) readings can arise for questions embedded under know. Following Groenendijk & Stokhof, George (2011) takes the SE reading as basic. Others, following Heim (1994), take the WE reading as basic, and derive more readings from it.

may yield extra constraints on possible strengthening mechanisms (for theories à la Heim, 1994) when hypotheses about the cognitive processes are made explicit.

In Experiment 1, we show that know can give rise not only to SE readings but also to weaker readings. In Experiment 2, using a different paradigm, we show that these weaker readings consist mostly of IE readings (and in fact both for know and predict). The results of Experiment 3 confirm that we cannot reliably establish the availability of the WE reading for know. Finally, in Experiment 4 we compare the processing properties of the IE and SE readings.

2 Experiment 1: Existence of WE/IE readings under know

2.1 Goal

Most theories derive various readings for sentences containing questions embedded under know (see Table 1). But most theorists report judgments that these configurations only give rise to SE readings. The goal of this experiment was to test this empirical claim. We did not aim to distinguish between WE and IE readings for the time being.

We also collected online data in order to test a prediction derived from Heim (1994). Her theory derives the SE reading from the WE reading. Hence, the SE reading requires an extra derivation step and we may thus be able to detect an extra processing cost.

2.2 Methods and Materials

The experiment consisted of a truth-value judgment task: Participants read pairs of sentences (which we call context and utterance, respectively) and had to judge whether the utterance was true or false in the given context. All experiments were conducted using American English.
2.2.1 Procedure

Participants were directed from Amazon Mechanical Turk to an online experiment hosted on Alex Drummond’s Ibex Farm. After filling out a consent form they received instructions (see details below), then completed a short training phase (8 items with feedback) followed by the experimental phase. After completing the experiment, participants filled in a questionnaire (including demographic questions about participants’ age, gender and native language). The final step was to validate their participation on Mechanical Turk.

2.2.2 Instructions

The instructions provided a story repeated in Fig. 1. The goal of this story was to provide a general context for the experiment such that we could easily manipulate what the truth-value of the various hypothesized readings would be. In a nutshell, it was set up so that various characters (e.g., the male baker) would have total (exhaustive) knowledge about some events (e.g., men buying bread) and partial or no knowledge about other events.

A small town in a remote area has two shopping malls: one for women and one for men.

• Men shop at the men’s mall, where all shopkeepers are men.
• Women shop at the women’s mall, where all shopkeepers are women.

In each mall there is one of each facility: there is one male and one female baker, one male and one female mechanic, and so on. Sometimes the shopping malls close.

• Men do not always know whether the women mall is open or not.
• Women do not always know whether the men mall is open or not.

Each shopkeeper knows what he/she sells to whom, but knows nothing about what happens in other shops.

Figure 1: The story given to participants in the instructions for Experiment 1

The instructions also included 4 example items, which are given in (10). (10a,c) are true and (10b,d) are false given the story in Figure 1. This was explained to the participants.9

(10)  
a. Situation: Today, the men shopping mall was open.
   “John-the-baker knows whether Bill bought bread.”
b. Situation: Today, both shopping malls were open.
   “Sandra-the-pharmacist knows who bought medicine.”

9There was a recurring typo in our materials: in the second part of the instructions and in context descriptions (not in target sentences) we dropped the genitive marker in (wo)men’s shopping mall. Thanks to an anonymous reviewer for catching this mistake.
c. **Situation:** Today the men shopping mall was closed and everybody, including the women, knows about this.
   “Lisa-the-tailor knows who bought a coat.”

d. **Situation:** Today, both shopping malls were open.
   “Henry-the-hairdresser knows if Peter got an oil change.”

### 2.2.3 Stimuli

The stimuli were built from 3 types of contexts, as exemplified in (11), and from 3 types of utterances, as exemplified in (12).

(11)  
   a. *Today both shopping malls were open.*
   b. *Today the women shopping mall was closed. Everybody knows about this.*
   c. *Today the women shopping mall was closed. Men do not know about this.*

(12)  
   a. John the baker knows who bought bread.
   b. John the baker knows which men bought bread.
   c. John the baker knows which men bought meat.

The crucial utterance was of the form of (12a). The truth value of its different possible readings varies with context as follows (see Table 2 for a summary):

- In context (11a), (12a) is false under any exhaustive reading because John does not know which women bought bread (assuming some women did, as soon as their mall is open). The items obtained from the combination of this utterance and this context were false controls.

- In context (11b), (12a) is true under all readings: For each man, John can tell whether he bought bread or not, and he knows that no woman bought bread that day. Therefore for each person John can tell whether he or she bought bread. The items obtained from the combination of this utterance and this context were true controls.

- In context (11c), (12a) is true under the WE reading because only men bought bread, therefore for each person who bought bread it is true that John knows that this person bought bread. It should also be true under the IE reading, because John has no reason to falsely believe that a given woman bought bread. However, it is false under the SE reading, because John does not know that no woman bought bread. The items obtained from the combination of this utterance and this context formed our targets. If only the SE reading is available, we shouldn’t expect differences between (12a) in contexts (11a) and (11c): both items would be false.

We created variations of each item by using the 16 characters presented in Table 3 (8 jobs \( \times \) 2 genders) and varying the words ‘men/women’ accordingly in the contexts and utterances. All 16 possible targets, along with true and false controls, were included. 24 true and 24 false fillers for which the truth value did not depend on the context were also created using
Table 2: The truth value of the crucial utterance (12a) under each reading depends on the context. Critically, answers in the (11c) context distinguish between the SE reading (false) and the other two (true).

Table 3: We used 16 characters. For each of the 8 types of shop, there was one male and one female name, and one matching and one mismatching activity. All activities were alternatively matching and mismatching, depending on the shop they were associated with.

2.2.4 Participants

40 participants were recruited on Mechanical Turk. 39 of them completed the task (17 females, 22 males). Their age ranged from 19 to 66 years (mean: 33). All of them reported English as their native language.

2.3 Results

2.3.1 Data treatment and Statistical methods

Responses made in less than 100ms or more than 10s were discarded (0.7% of the data). All mixed models were built with a maximal random effect structure based on participants and items as random variables (in the sense of Barr et al., 2013). Our item factor was a value between 1 and 8 corresponding to one line in Table 3. This means that two items involving
the same shop (e.g., ‘baker’) were not considered independent measures. For each model we give the estimate of the fixed effects (β) and a p-value with its associated statistics. For linear mixed models we used the t-statistics given in lme4 and for GLMM the χ²-statistics with one degree of freedom obtained by comparing the models with and without a given fixed effect.

2.3.2 Analysis of responses

Fig. 2a shows the proportions of True responses to targets, true controls, and false controls. Overall, the task was well-understood and executed as reflected by high accuracy on true (88%) and false (73%) controls.

On targets, participants gave on average 43% True responses and individual rates varied from 0% to 100% (while there was little between-item variation: 36% to 48%). Two logit mixed models were fitted to compare the responses to targets with true or false controls respectively. The results showed that the differences were significant between targets and both true controls (β = 2.7, χ²(1) = 32, p < .001) and false controls (β = −1.4, χ²(1) = 8.4, p = .003).

The fact that participants gave fewer True responses on targets than true controls suggests that some participants had a SE reading and treated the utterance as false on the target trials. More surprising is the fact that they treated targets differently from false controls. According to Table 2, this high rate of true responses reveals the presence of a reading other than the SE reading. The fact that the intermediate rate of true responses corresponded to individual rates from 0% to 100% confirms that participants differed in the way they understood the same sentence. If they were simply at chance because the target sentences were too complicated we might have observed a more homogeneous pattern around 50% across the participants; alternatively, some participants would have been at one end of the spectrum (those that managed to do the task) and others would have been at 50%.

2.3.3 Analysis of response times (no clear result)

On targets, False responses should reflect SE readings, while True responses should reflect WE or IE readings (and possibly some mention-some readings). We can thus compare the response times (henceforth: RT) associated with true and false responses, as indicative of how each reading is processed. We removed 8 participants whose error rate on controls and fillers was at least one standard deviation above the mean (threshold: 28%).

Alternatively, one may argue that some participants decided to answer true to complicated cases and others decided to answer false. If the reader is not convinced that we are seeing evidence for an ambiguity at this point, a potentially more convincing argument will be presented with the next set of experiments to be described.

These participants were not removed from the analysis of responses to avoid introducing distortions in this analysis. In short, to analyze responses we avoided removing participants based on their responses. More precisely, selecting only participants with a low rate of True responses on false controls but putting no constraint on their responses to targets could create an artificial difference between these two conditions that we eventually want to compare. Practically, we argue that the target condition shows high variability, but this can only be assessed against the controls and if the variability of the controls has been removed a priori, this comparison would make no sense. For the RT analysis, since we proved that there is a real
true controls  

(a) Percentage True to the different conditions. (Targets correspond here to WE+IE targets with the terminology of the following experiments.)

(b) Response times by answer. On targets both responses were acceptable (True = WE or IE, False = SE). On controls we kept only correct responses so True corresponds to true controls and False to false controls.

Figure 2: Truth value judgments for the target sentence in Experiment 1 and corresponding response times.

displays the response times of the remaining subjects. A model with Answer, Sentence type (Control vs. Target) and their interaction as fixed effects was fitted on the RT to controls and targets (after errors on controls were removed). None of the effects were significant (all $t$'s $< .5$). In subsequent experiments, we will use more powerful methods to investigate this issue.

2.4 Discussion

The offline results suggest that the SE reading is not the only available reading for questions embedded under know. This is surprising, since even theories which derive the WE reading as a primitive usually treat know as an exception which only gives rise to SE readings. The findings would thus make a strong point against theories from the second group in our classification (Groenendijk and Stokhof, 1982, 1984, 1993; George, 2011) since they can only derive SE readings.

Two alternative explanations of these results come to mind. First, one may wonder whether mention-some readings could be called for to explain the appearance of non-SE readings. Note however that mention-some readings should generate true responses in all conditions, including false controls. The rate of true responses to false controls thus provides

difference between these two conditions, we allow ourselves to consider True responses to false controls as errors while considering both True and False responses to targets as reflecting genuine linguistic judgments.
a baseline of errors and mention-some readings, but targets gave rise to significantly more true responses, which calls for an independent explanation. Second, the expert reader may wonder whether the relevant false responses are not in fact SE readings in disguise, e.g., as altered by so-called covert domain restriction. We will explain this possibility in further detail in section 3.5 and assess it in detail in Experiment 3. For the time being, we will proceed to confirm and refine the current set of results with a more efficient method, which will also help distinguish between two types of non-SE readings: WE and IE readings.

3 Experiment 2: non-SE readings for know and predict

3.1 Goals

Results from Experiment 1 provide reasons to believe that WE and/or IE readings are available for questions embedded under know. The main goal of Experiment 2 is to determine which readings participants have when they do not have the SE reading. In this experiment, we also compared different responsive predicates (know and predict). Klinedinst and Roth-schild (2011) in their actual implementation derive IE readings for predict but not for know, while Spector and Egré (2014) argue that both verbs can support the IE reading.

We also simplified our experimental paradigm into a picture-sentence matching task, which allows finer-grained distinctions between the three exhaustive readings while keeping the differences between the know and predict versions minimal. Arguably, it also simplifies the task in that the participants can proceed on an item-by-item basis and do not have to memorize a complicated story.

3.2 Methods and Materials

The experiment consisted of a picture-sentence matching task: each experimental item consisted of a sentence paired with a picture, as described in the Materials section and illustrated in (13) and (14). Participants had to judge whether the sentence was an accurate description of the picture. The recruitment procedure was similar to the first experiment, except that participants were randomly assigned to one of two versions of the experiment (involving either know or predict).

3.2.1 Instructions

The instructions included a short story, but it did not play a critical role. All the relevant information that was previously given in the story of Experiment 1 was now encoded in the pictures. Here is the story for the know-version of the experiment:

John is playing a card game to train his memory. Each turn, he picks a card, quickly looks at it and flips it. John tries to recall what he saw on the card. Then he looks at the card again, and checks if he was right. You will see the card, and how John remembered the card before checking it. Using this information you will have to judge whether some sentences about the card and John’s knowledge

12 A closer look at the data reveals that a few participants consistently answered True to false controls. These participants potentially had a mention-some reading.
are true or false. All cards are made up of 4 squares, each of which can be red, green or blue.

The story for predict was very similar, except that John was trying to predict the content of the card before looking at it.

The instructions also included 4 example items (2 true and 2 false) with some explanations on how to understand the pictures, as well as general instructions which were identical to those of the first experiment (e.g., about response buttons).

3.3 Stimuli

Examples (13) and (14) show items used in the know and predict versions of the experiment respectively.

(13) WE target (know-version):

(14) WE+IE target (predict-version):

3.3.1 Sentences

The sentences were variations on the sentence templates given in (15) and (16). Position was one of ‘top’, ‘bottom’, ‘left’ or ‘right’ and color one of ‘red’, ‘green’ or ‘blue’. The most important sentences were those built on (15a) and (16a).

(15) Sentences for know:

a. John knew which squares were [color].
b. The [position] squares were [color].
c. John knew which of the [position] squares were [color].

(16) Sentences for predict:

a. John predicted which squares would be [color].
b. The [position] squares were [color].
c. John predicted which of the [position] squares would be [color].

3.3.2 Pictures

As seen in examples (13) and (14), the pictures consisted of two parts. In both versions of the experiment, the left side of the pictures represented the actual card that John had picked. The right side of the picture represented John’s beliefs about the card, or the prediction
he made beforehand, depending on the version. This was explained in the instructions and repeated on each item. In some cases, such as (14), the right part (which represented John’s beliefs or predictions) included one or two gray squares with a question mark. This represented the fact that John sometimes had no idea about the color of these squares (know-version) or had not made any prediction about these squares (predict-version). This was explained in the instructions as well.

3.3.3 Targets

The two examples in (13) and (14) play a central role. As summarized in Table 4 and described below, together they allow us to infer which reading a given participant accessed (assuming participants were coherent):

- **WE targets**: Items like (13) are true under the WE reading and false under the stronger IE and SE readings. They are true under the WE reading because for all blue squares on the actual card, John knew that they were blue (respectively predicted that they would be blue). However, such items are false under the IE reading (and SE reading, consequently), because John had a false belief (respectively made a false prediction) about the bottom right square. Therefore, if a participant responds True to such items, it is presumably based on WE readings for the sentences.

- **WE+IE targets**: Items like (14) are true under the WE and the IE readings, but false under the SE reading. They are true under the WE and IE readings because for all blue squares on the actual card, John knew that they were blue (respectively predicted that they would be blue), and John did not falsely believe (respectively did not falsely predict) that any other square was blue. However, such items are false under the SE reading because John did not know for sure (respectively make an explicit prediction) that the bottom right square was not blue. Therefore, if a participant responds True to such items, it is presumably based on either the WE or the IE reading.

<table>
<thead>
<tr>
<th>Responses to</th>
<th>Inferred reading</th>
</tr>
</thead>
<tbody>
<tr>
<td>WE targets</td>
<td>WE+IE targets</td>
</tr>
<tr>
<td>True</td>
<td>True</td>
</tr>
<tr>
<td>False</td>
<td>True</td>
</tr>
<tr>
<td>False</td>
<td>False</td>
</tr>
<tr>
<td>True</td>
<td>False</td>
</tr>
</tbody>
</table>

Table 4: True responses to WE targets correspond to WE readings, while true responses to WE+IE targets correspond to WE or SE readings so that together responses to the targets distinguish among the three (WE, IE and SE) exhaustive readings.

3.3.4 Controls

In addition to these two types of targets, the experiment included a variety of controls. These controls ensured that (a) each sentence appeared with pictures that made it unambiguously
true or false, (b) each picture appeared with sentences that were accordingly unambiguously true or false, and (c) there was an equal balance of true and false controls, and a low proportion of targets (15%).

Of crucial importance were items which involved embedded questions, but were unambiguously true or false. They served as baselines for errors: they can help to prove that a certain category of responses may be rare but still more frequent than actual errors, and that therefore this category of responses corresponds to (rare but genuine) readings (see analyses below for details). Using the terminology above that encodes the set of true readings for a particular condition, these conditions were as follows:

- **∅ baseline**: false under each of the three exhaustive readings.
- **WE+IE+SE baseline**: true under each of the three exhaustive readings.

### 3.3.5 Item generation
The items were varied by rotating the cards in the pictures and adapting the [position] words in the sentences accordingly (4 possibilities), or by permutation of the colors in the pictures and in the sentences accordingly (3 possibilities). We also varied the amount of ignorance in WE+IE targets and corresponding controls (one or two gray squares).

Overall there were 240 items in each version of the experiment, including 12 WE targets, 24 WE+IE targets (8 with two gray squares, 16 with one). This gave a 15/43/43% proportion of targets/true controls/false controls respectively.

### 3.3.6 Participants
30 participants were recruited for each version, but one subject in the know-version did not complete the task. One more subject in the know-version was removed from the analyses because he reported being a native speaker of a language other than English. The age of the remaining participants ranged from 18 to 59 years (mean: 34), there were 31 females and 27 males. Eight more subjects were removed because their error rate on controls was more than one standard deviation above the mean (threshold for rejection: 12.1%).

### 3.4 Results

#### 3.4.1 Data treatment and Statistical methods
Responses made in less than 100ms or more than 10s were discarded (1.2% of the data).

All mixed models were run with maximal random effect structure when possible, but random slopes for subjects had to be dropped in the logit mixed models of section 3.4.2. When random slopes associated with the ‘baseline vs. target’ contrast for Subject were included, the predicted values differed greatly from the actual data, and the distributions of random effects for subjects were clearly not following a centered normal distribution. This seems to be related to a high proportion of subjects who gave exactly 100% True responses to WE+IE targets or 0% to WE targets. Dropping the random slope associated with Sentence Type for subject seemed to be the best workaround (Verb is a between-subject factor, hence there is no random effect associated with it).
but two items in different conditions under the same rotation were treated as dependent measurements.\textsuperscript{14}

For each model we give the estimate of the fixed effects ($\beta$) and $p$-values with the associated statistics. For all mixed models in this experiment, we used the $\chi^2$-statistics with one degree of freedom obtained by comparing the models with and without a given fixed effect.\textsuperscript{15}

### 3.4.2 Analysis of responses

The mean error rate on controls was 5.1%. This confirms that this task was easier than the one used in Experiment 1 (in which the error rate was 15%). No participant had more than a 7% difference in performance between any two different colors and the effects of color on global performance were not significant. If some participants had been red-green colorblind, we would likely have observed bigger differences.

Fig. 3a shows the proportion of True responses given to all types of items (baselines and targets). First, we compared the WE+IE targets with the WE+IE+SE baseline. These items only differ with regards to the SE reading, which is false on the WE+IE targets but true on the WE+IE+SE baseline. Only participants who accessed SE readings could treat these items differently, while participants with IE or WE readings should answer True in both conditions. A logit mixed model (with only random intercepts for subjects, as explained above) was fit on responses to the WE+IE+SE baseline and WE+IE targets with Verb (\textit{know} vs. \textit{predict}), Sentence Type (Baseline vs. Target) and their interaction as fixed effects. There were slightly more False responses to targets than to the baseline ($\beta = -0.7, \chi^2 = 4.7, p = .03$), suggesting that the SE reading was available. This effect interacted with verb: \textit{predict} gave rise to more False responses on WE+IE targets ($\beta = -1.4, \chi^2 = 11, p = < .001$), suggesting that if anything \textit{predict} gives rise to more SE readings than \textit{know}.

Second, Fig. 3a reveals that many more true responses were given to WE+IE targets than to WE targets, indicative of the important impact of the IE reading which distinguishes between these two conditions. A logit mixed model confirmed the strong difference between the two types of targets ($\beta = -7.2, \chi^2 = 59, p < .001$). The interaction with verb was not significant.

Finally, we tested for the presence of the WE reading by comparing WE targets with the $\emptyset$ baselines. These two conditions only differ in that the targets are true under the WE reading, while the baselines are not. A WE reading would thus lead to respond differently in these two conditions, while other readings should lead to False responses in both conditions. A mixed model was fit on responses to $\emptyset$ baseline and WE targets with Verb (\textit{know} vs. \textit{predict}), Sentence Type (Baseline vs. Target) and their interaction as fixed effects. We found a significant difference between WE targets and the baseline ($\beta = 2.4, \chi^2 = 39, p < .001$), suggesting that the WE reading was available. This effect interacted with verb:

\textsuperscript{14}Note that for controls which referred to a particular row or column in the card (e.g., ‘the top squares’), rotation also encoded differences in the sentence (the position words ‘top’, ‘right’, ‘bottom’, ‘left’). We may expect differences in decoding pictures where the crucial information is aligned vertically or horizontally, as well as differences due to position words (e.g., ‘top’ vs. ‘bottom’ may be easier than ‘left’ vs. ‘right’).

\textsuperscript{15}For models with more than one fixed effect, the $\chi^2$ for the simple effects corresponds to the comparison between models restricted to the intercept levels of other effects.
predict gave rise to fewer True responses than know on WE targets ($\beta = -1.0$, $\chi^2 = 11$, $p < .001$).

Fig. 3b presents the responses of each participant combining the two types of targets. It shows that the few False responses to WE+IE targets come from a small group of participants who consistently accessed an SE reading (bottom-left corner), and the few True responses to WE targets from a group of participants who consistently accessed the WE reading (top-right corner). The rest of the participants mostly answered true to WE+IE targets and false to WE targets consistently (top-left corner), indicative of their accessing an IE reading.

![Figure 3: Average and individual responses in Experiment 2. The three exhaustive readings appear to be present.](image)

3.4.3 Analysis of response times

As shown by the analysis of responses, there were very few SE readings in the population. We therefore focused our response time analysis on WE and IE readings, and removed 4 participants who possibly accessed SE readings (criterion: at least 50% False responses to WE+IE targets). We categorized 37 of the remaining participants as IE (at least 50% False responses to WE targets) and 9 as WE (at least 50% True responses to WE targets). We then removed, as errors, responses to WE targets which did not match the category assigned
As a result, we sought to analyze false responses for IE participants and true responses for WE participants. To control for the influence of true/false response effects on RT, we included in the model correct responses to conditions in which we expected responses that would match the to-be-analyzed responses for each participant: false responses to the ∅ baseline for IE participants and true responses to the WE+IE+SE baseline for WE participants. Fig. 4 shows RT for each category of participants.

We ran a linear mixed model on responses to WE targets and the appropriate controls (∅ baseline for IE participants, WE+IE+SE baseline for WE participants) with Subject category (WE vs. IE), Sentence type (Target vs. Control) and interaction as fixed effects and a maximal random structure for items and subjects.\(^{16}\) We observed a significant interaction between Subject category and Sentence type \((\beta = -0.3, \chi^2 = 14, p < .001)\), suggesting that WE responses were slower than IE responses, even after controlling for true/false biases.

Figure 4: Experiment 2, RT on WE targets and baselines by category of subject. IE subjects were faster than WE subjects on targets. This cannot be an effect of response as the two groups did not differ on controls of matched truth value.

3.5 Discussion

Discussion of the results. We observed little difference between know and predict. We confirmed that know supports readings other than the SE reading. IE readings were most visibly available, but some participants seem to have accessed the WE reading. We note

\(^{16}\)We first ran a linear mixed model with Subject category (WE vs. IE), Sentence type (Target vs. Control) and Verb (know vs. predict) and all interactions as fixed effects. We dropped the random effects associated with verbs because the maximal model did not converge. Even then, none of the fixed effects associated with Verb was significant, and a model without these effects was not less powerful \((\chi^2(4) = 4.4, p = .35)\), so we report the model as above.
however that this WE-like behavior could also have emerged from another phenomenon, namely covert domain restriction, as we will now explain.  

**Covert domain restriction.** Covert domain restriction is a general linguistic phenomenon arising as soon as some form of quantification is involved. As an example, if we look at the dialogue in (17), we understand that B’s response in (17a) is not about all the students in the world but only those students who attended the party.

Explicit restrictors usually inhibit covert restriction (see Westerståhl, 1985). For instance, adding the restrictor ‘in Europe’ to B’s response as in (17b) feels awkward as it is now nearly impossible to reinterpret it as “the students who attended the party” (even if the party took place in Europe).

(17) A: “How was the party on Saturday?”
   a. B: “All the students were very happy.”
      *Possible interpretation:* All the students who attended the party were very happy.
   b. B: “All the students in Europe were very happy.”
      *Unavailable interpretation:* All the students (in Europe) who attended the party were very happy.

Now we note that, quite generally, all three exhaustive readings become equivalent if we restrict the domain of quantification to individuals who satisfy the embedded predicate (in Experiment 2, *being blue* or for our introductory example (1), *being a caller*). Abstractly, this is so because the exhaustive readings only differ on the status of individuals which do not satisfy the predicate (i.e. non-blue squares, or non-callers), with the WE reading being the one that puts no constraint on these individuals.

Hence, if we restrict attention to individuals who do satisfy the predicate, then the three readings collapse: all exhaustive readings become equivalent if the domain is restricted to individuals who do satisfy the predicate. Concretely, all exhaustive readings of (18) can be paraphrased as in (19). Furthermore, these three exhaustive readings under domain restriction end up being equivalent to the WE reading of (20) without domain restriction. Hence, what may look like WE readings of (20) without domain restriction may in fact be any of the exhaustive readings with an underlying domain restriction.

(18) John knew which of the actual blue squares were blue.
(19) For each blue square, John knew that it was blue.
(20) John knew which of the squares were blue.

---

17 Several participants at the Euro-XPrag 2013 conference suggested another source for the appearance of responses seemingly generated by WE readings. There is a parallel between our task and the classical ‘Wason selection task’ (Wason, 1966): verifying that John is right only about the actual blue squares is equivalent to searching for positive evidence only in Wason’s task. We do not investigate this interesting issue further because the results of the next experiment will receive a straightforward interpretation in terms of domain restriction but not in terms of restriction to positive evidence. In fact, domain restriction and restriction to positive evidence have similar practical consequences, but may be generated by different pressures.
Therefore, as we explained, what we considered WE readings of (20) in Experiment 2 may in fact be IE or SE readings with covert domain restriction. In the next experiment, we will introduce explicit restrictors to inhibit covert domain restriction (as discussed in (17); see Chemla, 2009 for a similar move). Consequently, most of the alleged WE readings will disappear, suggesting that they were in fact due to covert domain restriction.

4 Experiment 3: WE reading or domain restriction?

4.1 Goal
The goal of this experiment was to test whether the WE readings we observed in the previous experiment were due to implicit domain restriction. To reduce the chances that participants would apply a covert restriction of the domain, we used an explicit domain restriction. Apart from this change, the experiment was fully identical to the know-version of Experiment 2. We decided to focus on one verb for simplicity and chose know because it gave rise to the highest rates of WE responses.

4.2 Methods and Materials
4.2.1 Materials
The task and materials were exactly the same as in Experiment 2, except for sentences such as “John knew which squares were blue” in which an explicit domain restriction was added, as shown in (21). If explicit restrictors block covert restriction, participants should only answer True to the new WE targets if they access a genuine WE reading. Therefore, if the True responses in the previous experiment corresponded to WE readings, adding an explicit restrictor should have no effect. On the contrary, if at least some of these responses corresponded to covert domain restriction, then we predict less True responses in this experiment.

(21) John knew which of the four squares were blue.

4.2.2 Participants
60 subjects were recruited on Mechanical Turk and completed the task. One was removed from the analysis for reporting to be a native speaker of a language other than English. One more participant was removed from the analyses because his behavior on targets was inconsistent (see Fig. 5b, participant at the bottom right). The age of the remaining participants ranged from 19 to 64 years (mean: 40) and there were 24 females and 34 males. 6 participants were removed because their error rate on controls was more than one standard deviation above the mean (the threshold was 10.7%, but all removed participants were above 18%).

—One may wonder about the possible import of processing times at this point. Two studies (Chambers et al., 2002; Schwarz, 2012) suggest that, at least under some conditions, covert domain restriction is fast, while our suspect responses are not and may thus not be based on domain restriction. But if our suspect responses are WE readings, it would also be natural to expect them to be fast, given that they play the role of primitives in the relevant accounts. Hence, at this point, the RTs are surprising either way and do not provide a reliable source of information about the nature of these responses.

21
4.3 Results

4.3.1 Data treatment and Statistical methods

Responses made in less than 100ms or more than 10s were discarded (1.5% of the data). The statistical methods are identical to those of Experiment 2. In particular, we also had to drop random slopes associated with sentence type (target/baseline) for subjects in GLMM.

4.3.2 Analysis of responses

Fig. 5a shows the rates of True responses in the different conditions. It also repeats the results from Experiment 2 for comparison. Fig. 5b shows average responses to both types of targets for each participant in Experiment 3.

As in Experiment 2, a logit mixed model was fitted on responses to the WE+IE+SE baseline and WE+IE targets with Sentence Type (Baseline vs. Target) as a fixed effect. The effect of Sentence Type was significant: Participants gave fewer True responses to WE+IE targets than to the WE+IE+SE baseline ($\beta = -1.7, \chi^2 = 43, p < .001$).

A similar model was fitted on responses to the $\emptyset$ baseline and WE targets. The effect of Sentence Type was significant: participants gave more True responses to WE targets than to the baseline ($\beta = 1.3, \chi^2 = 16, p < .001$).

4.3.3 Comparison with Experiment 2

Figure 5: Average responses in Experiments 2 and 3, and individual responses in Experiment 3. The availability of the WE reading is much reduced in comparison to Experiment 2, as shown by the difference in responses to WE targets (Fig. 2a).

As one can see in Fig. 5a, the rate of True responses to WE targets seems lower in this experiment than in Experiment 2, whereas the $\emptyset$ baseline does not seem to be affected. This
was confirmed by a GLMM on responses to WE targets and \( \emptyset \) baselines from Experiments 2 and 3 with Experiment (2 vs. 3), Sentence type (target vs. baseline) and their interaction as fixed effects. The interaction was significant: participants gave similar rates of True responses to \( \emptyset \) baselines in both experiments, but they gave less True responses to WE targets in Experiment 3 \((\beta = 0.8, \chi^2 = 13.5, p < .001)\).^{19}

4.4 Discussion

Although we still observed a difference between WE targets and the \( \emptyset \) baseline, the addition of an explicit restriction greatly reduced it. This confirms that the explicit restrictors we provided inhibit domain restriction (at least to some extent) and that what could have been interpreted as WE reading-based responses in Experiment 2 were probably due to SE/IE readings with covert domain restriction.\(^{20}\)

The residual True responses to WE targets are rare. They may be based on genuine WE readings or some remaining possibility for covert restrictions.\(^{21}\) In our last experiment, we will focus on the comparison between the IE and SE readings.

5 Experiment 4: IE vs. SE readings of know

5.1 Goal

The existence of IE readings for know being established, we wanted to compare the processing properties of the IE and SE readings. Therefore we focused on the acquisition of reliable online data about the two readings. This goal was achieved using a design with training. We ignored WE readings because they proved to be rare and hard to identify as such (see previous discussion).

5.2 Methods and Materials

5.2.1 Course of the experiment

This experiment was very similar to Experiment 2 (and, therefore, to Experiment 3 as well). The main difference concerned the division of the participants into two groups, dubbed “SE-group” and “IE-group”.\(^{17}\) The training session was longer and the corrective feedback during training was more detailed and adapted to each group. It was designed to bias participants toward the SE reading or the IE reading.

Bott and Noveck (2004) used a similar method in one of their experiments on scalar implicatures. The main advantage is to provide a 50/50 share of responses corresponding to each reading, and thus better online data. There are some differences between our design and that of Bott and Noveck (2004). First, we did not have each participant take the experiment with both types of training, so the training is a between-subject factor in our experiment.

---

\(^{19}\) The same model restricted to the data for know in Experiment 2 gave even stronger results.

\(^{20}\) One reviewer questioned this hypothesis and mentioned unpublished experimental data revealing an opposite effect with explicit restriction facilitating domain restriction. In our specific control experiment, explicit restriction did not facilitate domain restriction, because as expected the number of True responses on WE targets decreased. The circumstances under which explicit restriction facilitates or inhibits domain restriction would be a topic for closer scrutiny.

\(^{21}\) They may also be Wason type errors; see footnote 17.
Second, we used the same instructions for both training groups (the same as in previous experiments with know). The only difference between the two groups was implemented in the feedback they received during the training phase.

5.2.2 Materials

As mentioned above, the instructions were identical to the previous experiment. The training session was longer (28 items), so that subjects would have enough time to get the intended reading, and it now included 6 WE+IE targets. See (22) for an example of a training target and (23) for an example of a WE+IE+SE item (true baseline), both with their associated feedback. The feedback was adapted to the training group. On controls it was triggered by errors, and on WE+IE targets it was triggered by responses which did not match the intended reading: subjects from the SE group received feedback when they answered True, whereas subjects from the IE group received feedback when they answered False. The training phase also included four WE targets (see 24), on which we trained participants to answer false to fully prevent WE readings. The feedback was always a paraphrase of the reading we wanted the participants to access.

(22) WE+IE target (boldface added to stress differences between the IE and SE feedbacks)

<table>
<thead>
<tr>
<th>The actual card</th>
<th>John's beliefs</th>
</tr>
</thead>
<tbody>
<tr>
<td>“John knew which squares were blue”</td>
<td></td>
</tr>
</tbody>
</table>

a. SE feedback: It was false. The top squares were blue and he knew it. The bottom squares were not blue but he didn’t remember whether they were blue or not.

b. IE feedback: It was true. The top squares were blue and he knew it. The bottom squares were not blue and he didn’t remember them as blue.

(23) WE+IE+SE baseline item

<table>
<thead>
<tr>
<th>The actual card</th>
<th>John's beliefs</th>
</tr>
</thead>
<tbody>
<tr>
<td>“John knew which squares were blue”</td>
<td></td>
</tr>
</tbody>
</table>

a. SE feedback: It was true. The top squares were blue and he knew it. The bottom squares were not blue and he knew they were not.

b. IE feedback: It was true. The top squares were blue and he knew it. The bottom squares were not blue nor did he remember them as blue.

(24) WE target (treated as false controls in this experiment)
a. SE feedback: It was false. John knew that the top squares were blue. He didn't know that the bottom right square was not blue.

b. IE feedback: It was false. John knew that the top squares were blue. He remembered the bottom right square as blue when in fact it was red.

5.2.3 Participants

Two groups of 40 subjects were recruited on Mechanical Turk. One participant did not complete the experiment and one was removed for not being a native speaker of English. Remaining subjects ranged from 18 to 66 years (mean 32). There were 38 females and 40 males.

For the response time analyses, 8 participants were removed because their error rate on controls was more than one standard deviation above the mean (threshold: 13.4%). We also removed 6 participants whose error rate on targets (with respect to the expected reading for their group) was at least one standard deviation above the mean (threshold: 36.3%). Finally, 2 subjects were removed because they had at least twice more True responses to WE targets than errors on controls; therefore we suspected they accessed a WE reading (or covert domain restriction).

5.3 Results

5.3.1 Data treatment and statistical methods

As in previous experiments, responses made in less than 100ms or more than 10s were discarded (1.9% of the data).

For response time analyses, we only looked at control sentences which did not involve embedded questions (i.e. generated on the template in (15b)). Indeed, on sentences which embed questions, SE and IE subjects should access different readings. Hence if there is a processing difficulty associated with one reading, it should affect these controls too. Controls with simple sentences on the other hand should not be affected by the training.

Categorical True/False responses were not thoroughly analyzed, relevant differences were simply assessed with $t$-tests on arcsine-transformed proportions of True responses.

Mixed models on response times were fitted after log transformation, to comply with the homoscedasticity hypothesis. We give the $t$-statistics and the $p$-values obtained by treating these $t$’s as $z$-statistics (this is a reasonable approximation with 40 subjects per condition).

5.3.2 Analysis of responses

The training included 6 WE+IE targets. Fig. 6a shows the responses of subjects from both groups on these targets, as a function of their order of presentation during the training phase. Although participants from both groups started with a high rate of True responses
(they likely accessed an IE reading), one can see that participants in the SE group quickly switched to False responses, therefore converging on the intended SE reading. Interestingly, participants in the SE group gave significantly fewer True responses than participants in the IE group even on the very first target of the training phase (76% vs. 97%, $t(42) = 2.6$, $p = .012$). This reveals that a quarter of the participants in the SE group switched to the SE reading just by reading the appropriate paraphrase in the feedback of control items, even though the truth value of the SE reading was irrelevant for these items. In short, these participants did not need the feedback on WE+IE targets which explicitly associated these items with a False response.\(^{22}\)

Fig. 6b shows the average responses to WE and WE+IE targets in the experimental phase for each participant (including participants with some WE responses or high error rate on targets who were removed from the RT analyses). Most participants fall in the category corresponding to their group and no participant displayed a predominant WE pattern.

\[\text{(a) Training phase: Percent True responses on WE+IE targets from the training phase as a function of their order of appearance.}\]

\[\text{(b) Experimental phase: Proportion True responses to targets (see Fig. 3b for details).}\]

Figure 6: Experiment 4: Most participants started with the IE reading, but participants in the SE-group quickly converged on the SE reading. In the experimental phase, most participants fell in the intended category.

\(^{22}\)Participants in the SE group saw 3 control items (median value) before the first target, and 13 of the 34 SE-participants (38%) made at least one error before encountering their first target. We observed a trend for a correlation between the number of errors made before the first target and the chances of getting the SE reading without explicit feedback: each error on previous controls increased the chance of giving a False response to the first target by 18% ($t(32) = 1.8$, $p = .09$).
5.3.3 Response times

Response times in a picture-sentence matching task can be affected by features of both the pictures and the sentences, as well as interactions between them. For instance, a sentence such as “John knew which of the top squares were blue” will probably draw more attention to the top row in the picture. Pictures in which John was ignorant about one square could have been read differently depending on the sentence associated with them, but also depending on the training. For instance, it may be easy for a participant of the SE group to answer False when there is a question mark in the display.

Fig. 7 shows response times to targets and different types of control items which did not involve embedded questions. We separated control sentences by truth value and by presence/absence of a question mark in the picture.

We ran a mixed model on control sentences with truth value (true vs. false), presence or absence of a “?” in the display, training (IE vs. SE) and all possible interactions as fixed effects. The random effect structure was maximal for subjects and items. No effect reached significance (all $t'$s $< 1.5$). If anything, there was a trend for a main effect of truth value (false sentences taking longer than true sentences) and an interaction between the presence of “?” and the training.

We ran a second mixed model in order to compare responses to WE+IE targets with responses to controls. Since targets always had a “?” (a crucial feature to distinguish SE and IE readings), we compared them with controls that also had a “?”). The fixed effects were Answer (True vs. False)$^{23}$, Sentence type (control vs. target) and their interaction. Both main effects came out insignificant ($t < 1$), but the interaction was significant ($\beta = .12$, $t = 2.7$, $p = .007$). Targets incurred an extra delay for SE subjects, and this cannot be just an artifact of the presence of “?” in the display (otherwise the effect would not interact with the content of the sentences).

5.4 Discussion

We saw that the training had a very strong effect on participants in the SE group, who quickly converged on SE behavior. One may worry that the participants’ responses in this experiment reflect some artifact induced by the training procedure more than genuine linguistic judgments revealing the existence of SE readings. Let us provide two arguments to alleviate such a worry before discussing the consequences of our results. First, we can show that participants in the SE group did not generalize the explicit training on WE+IE targets (‘respond False’) to other experimental items with similar displays. One specific and possibly salient feature of the display in the WE+IE condition is the presence of a question mark, but the SE participants accurately responded True on true controls where the display contained a question mark (94%), showing that participants did not simply respond False based on the mere presence of this visual cue. Second, on training items which were not

---

$^{23}$On controls, the answer parallels the truth value of the sentence. On targets it parallels the training of the subject ($\text{True} = \text{IE}$, $\text{False} = \text{SE}$). Therefore, Answer is sufficient to encode the training. Since Answer is only used to encode well defined experimental conditions here, it does not fall into the category of “bad controls” for a regression on RT.
WE+IE targets participants received feedback which consisted of paraphrases of the reading we intended to convey to their group even though the expected response was the same in both groups. We saw that a significant proportion of the participants adopted an SE pattern of behavior before being told explicitly to treat WE+IE targets as false items, thus suggesting that the paraphrases played an important role in the training. This would not be the case if participants’ responses were only due to association between a feature of the display and a given response.

Overall, the offline data from this experiment (training and experimental phases) confirm that the IE reading is very salient in this paradigm, but the fact that a quarter of the participants in the SE group switched to the SE reading before receiving explicit feedback suggests that the SE reading is naturally available as well.

The online data indicate that SE responses took longer than IE responses, after controlling for effects of the display and truth value. A first interpretation is that the derivation of the SE reading is more complex, or even that it is derived from the IE reading. However the sustainability of such a conclusion is weakened by the asymmetry between the salience of the two readings in this paradigm. In fact, an alternative interpretation of the online results would be that they simply reflect the salience of the IE reading and an effect of training. Participants in the IE group may not even be aware of the SE reading. Participants in the SE group, on the other hand, must be aware of both readings since they started the training with the IE reading and switched to the SE reading. Therefore the extra processing cost for the latter group may reside in an ambiguity made salient for them between a true and
Let us note however that if we take the online data to reflect the salience of the IE reading, it strengthens the main conclusion we want to draw from the offline data, namely that the IE reading is naturally available.

6 General discussion

6.1 Summary of the results

In this series of experiments we showed that the SE reading is not the only reading available for know (Experiment 1). It seems that the most prominent reading in this task is the intermediate IE reading proposed by Spector (2005) (Experiments 2, 3, 4). Overall, we found fairly small traces of the WE reading, and we showed how this evidence was polluted by the possibility of covert domain restriction (Experiment 3).

The relative proportions of the different readings differed slightly for the verbs predict and know, but the general pattern was the same. Crucially, the IE reading was available for both verbs (Experiment 2).

Finally, the online results indicate that the IE reading is faster to access than the SE reading in our task. At the very least, this confirms the existence of the IE reading. This processing asymmetry may merely reflect the offline preference for the IE reading in our paradigm, which would then count as a confirmation of the conclusions obtained from the offline results. But, more tentatively, it may also serve as the basis for developing theoretical models informed by processing results, such that the operations leading to the IE readings could be proved to be simpler than (or a subset of) the operations leading to the SE readings.

6.2 Consequences for the theories

The availability of exhaustive readings beyond the SE reading contradicts the predictions of theories we categorized in the second group (see section 1.4(a)). These correspond to theories that rely on strong answers as the basic denotation of a question and can only derive strong readings for embedded questions (Groenendijk and Stokhof, 1982, 1984, 1993; George, 2011). On the other hand, theories from the first group, relying on weak answers derive the WE readings as basic and may, depending on the specifics of how the other parameters are set, derive IE and SE readings (Heim, 1994; Beck and Rullmann, 1999; Guerzoni and Sharvit, 2007; Klinedinst and Rothschild, 2011; Spector and Egré, 2014). Currently, the only two theories which explicitly implement IE readings, which were found to be salient, are those by Spector and Egré (2014) and Klinedinst and Rothschild (2011). In both cases, the way to obtain these IE readings for factive verbs such as know is to rely on a separation between the assertive and presuppositional components of the verb (Spector and Egré implement this separation between assertion and presupposition explicitly, while Klinedinst and Rothschild only provide guidelines on how this could be done).

This hypothesis is consistent with a wide range of processing approaches, including sequential sampling process models of decision making (e.g. Ratcliff et al., 2004; Ratcliff, 2002, 1979; Tuerlinckx, 2004; Busemeyer and Rapoport, 1988; Vickers, 1979; Pike, 1973, 1968, 1966); it may be directly linked to classical results in cognitive psychology which show that when distinct aspects of a given stimulus push subjects in different directions, their decision process takes longer, regardless of the actual outcome.
The current results thus provide constraints for future theories of embedded questions. They suggest that (1) weak answers are primitives and (2) IE readings exist. This latter result argues for the possibility of including in the semantic system the tools to decouple presuppositional and assertive content.

To conclude, let us mention two potential questions for future research. First, question-embedding predicates can vary along many dimensions (e.g., presuppositionality but also polarity). Such properties play a different role in current theories of embedded questions, such that one should next test a wider variety of verbs using comparable methods. Such data would inform us on the plausible mechanisms for combination between a responsive verb and a question (see section 1.4(b)). Second, we have reported on the first processing data concerning embedded questions. But at this point it is hard to determine completely what our data reflect and what theories actually predict in terms of processing. Further processing data and explicit implementations of the different theories provide a direction for future research. We hope to have shown that these are not out of reach.

References


