On the (non-)exhaustivity of naturally occurring *wh*-questions

Morgan Moyer & Judith Degen
Université Paris-Cité, Stanford University

October 20, 2022

Abstract
Blah blah blah

1 Introduction

A root *wh*-question can be answered in multiple ways.

(1) a. Where can I find coffee?
b. Who came to the party?

The most natural way to answer (1a) is to mention a nearby coffee shop, while the most natural way to answer (1b) is to provide an exhaustive list of party-goers. That is, out of the blue, (1a) receives a *non-exhaustive* interpretation, while (1b) receives an *exhaustive* one. We refer to these readings as 'Mention-Some' (MS) and 'Mention-All' (MA), respectively (following Hintikka, 1976; Groenendijk & Stokhof, 1982, 1984). MS and MA are also available as possible readings of embedded *wh*-questions:

(2) a. Scully knows where I can find coffee.
b. Scully knows who came to the party.

(2a) is true if Scully knows an MS answer to the root question (1a), while (2b) is true if Scully knows an MA answer to the root question (1b). On the surface, *wh*-questions do not specify whether they are intended to be MS or MA. What makes (1a) and (2a) naturally MS, and (1b) MA? In this research, we offer a contribution to answering that question by looking at the distribution of MS and MA interpreatations in naturally-occurring *wh*-questions.

We will use the term 'interpretation' to refer to the final MS or MA reading that a hearer lands upon in a context. The MS interpretation in (1a) is licensed by contextual goals (Groenendijk & Stokhof, 1982, 1984): if (1a) is asked by a tourist whose goal is to drink a coffee, then an MS interpretation is more felicitous; if asked by a coffee distributor whose goal is to explore the local market, then an MA interpretation is more felicitous. However, not all *wh*-questions lend themselves to MS, while MA appears to be available for most

(if not all) *wh*-questions . Indeed, (1b) is often presented as a paradigm *wh*-question that does not allow MS.

Previous work on the availability of MS/MA interpretations of *wh*-questions has typically focused on a small number of examples passed down through the literature, on the implicit assumption that *wh*-questions do not vary greatly. () Recent work the availability of MS/MA experimentally has systematically varied both linguistic and contextual factors across a broader range of sentences and contexts, with the goal of creating a more stable empirical basis for a theory of *wh*-question meaning (Moyer & Syrett, 2019; Moyer, 2020). That work has revealed that the availability of MS/MA is a function of both linguistic form of the questions and the speaker's contextual goal. On the one hand, while questions without modals are on average less acceptable on MS readings, explicit non-exhaustive discourse goals (for example, searching for a cup of coffee as compared to searching for people trapped in a burning building) can actually bolster the acceptability of a MS interpretation of such non-modal questions. On the other hand, MA interpretations can be equally *infelicitous* when discourse goals are non-exhaustive. The flexibility of interpretation in both directions has often been overlooked in the literature, with the exception of Ginzburg (1995); Asher and Lascarides (1998); Beck and Rullmann (1999); Lahiri (2002). In part, this could be due to the focus on a select few of examples, and the methodological attempt to separate out the effect of context. We discuss this methodological point in section XXX.

In the next section, we will discuss both the main theoretical treatments of Mention-Some, and review the factors which have been observed to affect the distribution of this reading, the presence/absence of modals, the wh-word, and matrix embedding verbs for embedded questions. We will then summarize with the main predictions derived theoretically and from observation of the literature.

In the current study, we test which claims about MS/MA licensing previously made on the basis of a few hand-selected or artificially constructed examples generalize to thousands of naturally occurring *wh*-questions . We operationalize the availability of MS and MA interpretations via the acceptability of question paraphrases elicited in a paraphrase rating task. Experiments 1 tested root (Experiment 1a) and embedded (Experiment 1b) *wh*-questions in their immediately preceding discourse context (Section XX). We find that some but not all predictions from the literature are borne out, and in particular we find no overwhelming MA bias, but we do find modulation of interpretation based on both Wh-word and modality, and matrix verb to some extent. In section XXX we present a second set of studies (Experiments 2a and 2b), using the same stimuli and task as the first two experiments. Crucially, in these studies we presented participants with target questions *without* their discourse contexts. Surprisingly, and in stark contrast to the literature predictions, we find significant bias for MS interpretations, although overall participants were more uncertain about the distributions of question meanings. As for our other factors of interest, we still find effects of wh-word and modality, although to a much lesser degree.

The picture that emerges is that there is considerable variation in *wh*-question interpretation. That variation is the result of both semantic (e.g., the presence/absence of existential modality, or the question-embedding verb in the case of embedded questions), and additional pragmatic factors (e.g., the resolution of the *wh*-domain, and the role of contextual discourse goals). We argue that MS is not an exceptional fringe phenomenon,

and that instead, an adequate theory of *wh*-question meaning will have to account for the interplay of both kinds of factors. On this basis of our data, we think that the maximal semantic claim that can be made is that questions are generally either ambiguous or underspecified for MS or MA, but that other factors like the presence of existential (modal) force, the pragmatics associated with resolving the *wh*-domain and reasoning under uncertainty about the speaker's goals, drive hearer perceptions about (non-)exhaustivity in *wh*-questions.

2 Mention-Some and the semantics/pragmatics of *wh*-questions

The theoretical landscape with regards to the semantics/pragmatics of *wh*-questions is highly disputed, and it raises theoretical and empirical issues that extend farther beyond the topic of *wh*-questions . There are consistent empirical observations which can and are explained in theoretically conflicting ways; even the data themselves are questioned. That makes a straghtforward disucssion–and theory testing–approach difficult, to say the least.

With that being said, we can say that there are two simple ways to approach MS: either it is semantic or it is pragmatic. Semantic treatments aruge that there is an underlying semantic explanation for MS, in other words that MS is grammatically licensed. In contrast, pragmatic accounts argue that pragmatic factors explain MS. This is a classic juxtaposition.

What are pragmatic factors? For Karttunen and Groenendijk & Stokhof, it means a non-truth conditional process that determines what counts as a good answer (see, Karttunen (1977) footnote 4; Groenendijk and Stokhof (1984) footnote 14 and discussion pp. 533). This understanding of "pragmatic" is maintained in George (2011) and indirectly in (Xiang, 2016) in the form of arguments against embedded "pragmatic" MS. The idea is that what counts as a good answer depends on the speaker/hearer interests, and in the embedded case, this information cannot penetrate grammatical structure. For Asher & Lascarides and Ginzburg, "pragmatic" implies information about a conversational participant's beliefs/cognitive state, and goals/plans (seee discussion on pp. 266 of Asher & Lascarides, 1998). The difference between these two perspectives is whether this pragmatic information has truth-conditional repurcussions; whether this information is available to the internal knowledge representations involved.

The problem here, is that drawing the line based on truth-conditional effects doesn't actaully aid either the empirical or theoretical situation. This is because of context-sensitivity.

Drawing a distinction in this way confounds two separate perspectives on language *understanding* and linguistic theory. Both present useful insights into how a *hearer* determines the meaning of a *wh*-question , but crucially make different assumptions about both theory and data. By introducing these perspectives, we will provide a frame in which to elaborate on our own view on MS and *wh*-questions .

In his 1992 book *Arenas of Language Use*, Clark described two approaches to linguistic inquiry: the language as *product* approach and the language as *action* approach.

Product traditions view the sentence as a basic unit, an abstract representation produced by a grammar. In the realm of semantics, a semantic theory follow the product tradition, abstracting away as much as possible from any circumstantial or indexical information. Groenendijk & Stokhoff allude to this in footnote 14 of the 1984 joint-dissertation, which describes their motivation for the semantics/pragmatics divide that they assume. Additionally, this thinking seems to underly George's rejection of a van Rooij style semantic theory, which supposedly gives up simplicity for incorporating more contextual/intentional/mental state information into the semantic theory (discussion on p.205). Often, such contextual information is acknowledged as playing a role in language production and comprehension, but then disregarded as playing a role in determining the truth (or answerhood) conditions of *wh*-questions . It has often gone hand-in-hand with the idea that pragmatic information cannot "interfere" with semantics, and thus the prediction that MS should not occur in (at least some) embedded questions–see also footnote 4 Karttunen (1977), and (Xiang, 2016)[mm: page number].

In contrast to product theories, action traditions follow in the steps of speech act theorists and ordinary language philosophers like Austin, Searle and Grice. These traditions hold utterances to be the more basic unit, and therefore equally important to the sentences which they token, are the contextual circumstances in which they are tokened. This includes indexical information about the speaker, hearer, goals. Utterance 'meaning' then refers to what Grice and Austin called 'speaker meaning', and is guided by the recognition of speaker and hearer intentions. It's possible that theories along the lines of Ginzburg and Asher & Lascarides would be candidates for being considered action theories, because their analyses of *wh*-questions are parts of the analyses of dialogue more generally, often appeal to situational information like cognitive states and plans.

In this work, we adopt the perspective of the langauge as action tradition. Our guiding inquiry, then, concerns the *interpretation* of *wh*-questions , which happens by necessity in fixed naturalistic contexts, with conversational agents present, and actively engaged (presumably) in cognitive processes like mind-reading/intention-recognition. We will treat semantic theories as presenting minimal models of the conditions on interpretation. The linking assumption is that, if a theory puts a mechanism in the semantics, the reason is because that mechanism is by hypothesis part of the basic meaning of the linguistic item in question. Thus, by convention, this predicted semantic component presumably should be borne out more often than not empirically—that is, in the interpretations that naïve language users access on average. The behavioral patterns that we see in naïve language users should reflect the presence of the hypothesized mechanism.

With these comments in mind, we turn to the behavioral patterns that the semantic theories above predict when it comes to *wh*-question interpretation.

In the following sections, we discuss the approaches to MS, and in the process identify four critical observations about factors that modulate the interpretation of MS, and *wh*-questions generally. These four factors have served differentially as evidence for semantic and pragmatic accounts of MS, thus our discussions of theory will be tied up with the evidence that has been produced in support of those theories. To preview, the effect of existential (modal) force, and matrix question-embedding verbs will unroll in the section on semantic theories of MS; while the effects of *wh*-word and contextual discourse goals will be discussed in the section on pragmatic approaches.

2.1 Semantic treatments of MS and predictions about the distribution of MS and MA

Many researchers have implicity assumed that MS is more limited in distribution than MA, assuming that the predominant reading/answer is MA in most cases. As a result, many have maintained theories which posit only an underlying MA semantic representation usually based off Karttunen (1977) or Groenendijk and Stokhof (1982, 1984). Some places where this view is explicitly stated (Groenendijk & Stokhof, 1982, 1984; Karttunen, 1977; van Rooij, 2003; Nicolae, 2014; Fox, 2014; Xiang, 2016; Dayal, 2016; George, 2011, Ch.6).[mm: xiang2020,fox2018].

Thus, if the bare semantic contribution of a *wh*-question is MA, then on average, we should find MA interpretations rated higher than MS interpretations. Additionally, we should find that MA interpretations will persist (or even increase) reqardless of whether contextual information is provided.

Prediction 1: General MA bias.
MS is limited in distribution, but MA is availble for all *wh*-questions .

Despite this often implicit assumption, there are those who argue for ambiguity in *wh*-questions . Theories like Hintikka (1974); S. R. Berman (1991); S. Berman (1994); Beck and Rullmann (1999); Lahiri (2002); George (2011, Ch. 2) [mm: make sure the IS citations are correct] (Ciardelli, Roelofsen, & Theiler, 2016; Ciardelli, Groenendijk, & Roelofsen, 2013; Theiler, 2014; Theiler, Roelofsen, & Aloni, 2016) posit that all questions in principle can allow for MS or MA.[1] For instance, Hintikka originally argued that questions were ambiguous between existential (MS) and universal (MA) readings. Beck and Rullmann (1999) aruge that there is a range of readings available for *wh*-questions and thus that the semantics should provide the tools to account for that. Their theory posits several different type-shifting answerhood operators (following (Heim, 1994)). They acknowledge that the final interpretation will be determined by the pragmatics of disambiguation. Lahiri (2002) agrees, though his operator includes (i) a context sensitive variable to allow for restrictions imposed by a matrix verb, and (ii) a covert quantifier with a meaning similar to *enough* whose strength varies contextually. George's baseline theory also allows for ambiguity (at least a semantic MS (via existential quantification) and a semantic (strong) MA), due to the presence/absence of an exhaustivity operator at LF. However, George seems to think that their theory overgenerates (and Fox (2014) agrees), and so they present a restricted version later in Chapter 6.

Restricted ambiguity theories like Fox (2014); Nicolae (2014); Xiang (2016); George (2011, Ch. 6)[mm: fox2018, xiang 2020] posit ambiguity in a subset of *wh*-questions . Consider the examples below.

(3) a. Scully knows where I **can** find coffee.

[1] Theories like (Ginzburg, 1995; Asher & Lascarides, 1998; van Rooij, 2003, 2004) also make this claim that in principle all questions should all MS and MA. We categorize these theories however as pragmatic because the explanatory brunt behind what makes a question MS or MA is going to crucially involve pragmatic information. At the same time, the same will be true of so-called semantic theories that posit ambiguity, because contextual factors are implicated in disambiguation.

b. Scully knows where **to find** coffee.
c. Scully knows where **a pen** is.
d. Scully knows who has **a light**.
e. Scully knows who some of the people at the party are.

What these examples have in common is that they involve an existential element: (a) and (b) have either an overt (a) or covert (b) modal (Bhatt, 1999; Dayal, 2016; Xiang, 2016; George, 2011; Nicolae, 2014; Fox, 2014); (c) and (d) have an existential indefinite (Groenendijk & Stokhof, 1982, 1984; van Rooij, 2003, 2004) [mm: chierchia 2013? roelofsen and IS people]; and (e) has an existential quantifier. Motivated by the observation that MS consistently correlates with the presence of some kind of existential item, these researchers have argued that only *wh*-questions with such elements are semantically ambiguous (Nicolae, 2014; Fox, 2014; Xiang, 2016; George, 2011, Ch:6)[mm: fox2018, xiang 2020]. Without going in to too much detail, the common theoretical explanation for why these kinds of *wh*-questions are ambiguous is because the existential element enters into a scope ambiguity with another element or operator in the question. Thus, our second prediction about the distribution of MS/MA is that modal questions will be MS biased, but non-modal questions will not be.

For these theories, the bare semantic contribution of a *wh*-question is MA, *unless* there is an existential element. Only in these modal or existential questions is there the possibility of a semantic MS.

Prediction 2: Modal MS bias.
Modal questions will give rise to MS on average more than non-modal questions.
Non-modal questions will show an MA bias.

[mm: Non-modal prediction is For Judith to smooth out so it doesn't sound like we're firing (too many) shots. tho it continues to be annoying going through all the quotes where these people EXPLICITLY say this, and then be told by them that I'm being unfair in my interpretation of the lit....] We've included in this prediction that non-modal questions should exhibit an MA bias. On the one hand, this prediction seems perfectly reasonable for semantic theories on which MS is semantic only in virtue of the modal, becuase these theories posit mandatory covert exhaustivity operators that presumably render question MA in the absense of the crucial existential modal, ceterus peribus. This interpretation of the literature has supported explicitly in several places ([mm: Nicolae (2013), p.17, Fox 2018 in section 4 (pdf p.12) and Xiang, under review paper, pp.16]). However, some have objected to this interpretation of these theories, suggesting that questions should be examined on a case-by-case-basis (Xiang, p.c, Spector p.c.).

Another point of variation predicted by some semantic theories involves the (semantic) selectional restrictions imposed by particular question-embedding verbs (following Grimshaw, 1979). The most frequently discussed are the restrictions imposed by the verb *know*, which many agree seems to select for MA ((Heim, 1994; George, 2011)[mm: zimmermann 2010]). However, Beck and Rullmann (1999); Sharvit (2002) argue that *know* allows for weak exhaustivity as well and ambiguity theories in principle are flexible enough to allow for the range of readings with *know* or other verbs.

These generalizations have more recently been called in to question both theoretically and experimentally (Cremers & Chemla, 2016, 2017; Klinedinst & Rothschild, 2011; Theiler, 2014) [mm: uegaki Sudo 2020]. Experimental work by Cremers and Chemla found non-strongly exhaustive readings with *know-who* questions (Cremers & Chemla, 2016), as well as strongly exhaustive readings with emotive factives like *surprise* (Cremers & Chemla, 2017). Note that this flexibility revealed experimentally is compatible with the flexibility we have already seen with the *know-wh* reports in (3a-c) which are acceptable on MS readings. However, these studies did not address the availability of MS. Experimental results looking specifically at MS, Moyer and Syrett (2019) did find differences between *know-wh* and *predict-wh*, where MS was more acceptable with the latter than the former. However, when discourse goals were systematically manipulated, even constructions with *know* were acceptable on MS readings.

Thus, we might predict differences with respect to difference matrix verbs, although it is an open question how strong these differences will be, and the extent to which they will be modulated by interactions with other factors, like the *wh*-word , modality, and contextual discourse goals.

If the semantic contribution of a *wh*-question embedded under *know* is MA, then presumably on average, hearers will interpret *know-wh* as MA on average more than MS, as well as on average more than when a *wh*-question is embedded under another verb.

Prediction 3: *know-wh* MA bias.
On average, *wh*-questions embedded under *know* will yield higher ratings for MA readings than for MS readings, and more than *wh*-questions embedded under other verbs.

2.2 Pragmatic treatments of mention-some

Pragmatic treatments of MS(or, *wh*-questions generally) argue that certain aspects of the context, like the speaker's interest/goal/plan and/or mental state play a crucial role in determining whether a question is interpreted MS or MA. Treatments that fit this criteria would include (Ginzburg, 1995; Asher & Lascarides, 1998; van Rooij, 2003, 2004). Theories which argue for an underlying MA semantics might fall into this category as well (Karttunen, 1977; Groenendijk & Stokhof, 1982, 1984), because they predict that only MS is sensitive to discourse goals in this way.

For example, in a now classic discussion, (Groenendijk & Stokhof, 1982, 1984) present (4) as evidence that MS requires special licensing from "background concerns," (Groenendijk & Stokhof, 1984, p.543).

(4) Where do they sell Italian newspapers in Amsterdam?

The scenario that first comes to mind is that of an Italian tourist looking for word of their homeland, in which case the MS interpretation is the most natural. The "background concern", or *goal*, of the questioner is to acquire the newspaper. However, the same question asked by a newspaper distributor interested in learning about the possible locations to distribute Italian newspapers for profit makes MA most natural. G&S would arugment that,

while the semantic contribution of the question is MA, the background concerns permit the acceptability of an MS answer only *in an MS context*.

Other researchers have taken such data as evidence for different analyses of the underlying semnatic representation, presenting semantic theories that include variables parameterized to aspects of context (Boër & Lycan, 1975; Ginzburg, 1995; Asher & Lascarides, 1998; Lahiri, 2002; van Rooij, 2003, 2004)[mm: boer and lycan 1985]. The resolution of these variables crucially involves fixing certain apects of the speaker's beliefs and goals/plans. To prime the intuition, Asher & Lascarides consider the following example:

(5) a. Mulder: How do I get to the burried treasure?
b. Scully: You go to the secret island.
c. Mulder: Scully knows/told me how to get to the burried treasure.

Regardless of (b) being an answer to (a) in the standard sense, whether or not Mulder's statement in (c) is true will *also* depend on what his goals are, and whether he knows first how to get to the secret island. If he needs to get to the burried treasure but does not have that prerequisite knowledge, then (c) is judged to be false. Thus, even the truth of (c) crucially depends on the speaker's plan and cognitive state.

Thus, if the interpretation of a *wh*-question (MS or MA) depends on this information, hearers' judgements interpretation of a *wh*-question will depend on whether the interpretation aligns with the perceived speaker goal. Additionally, if contextual information is removed, then we should see more uncertainty in hearers' judgements.

Prediction 4: Goal-Sensitivity. On average, MS should be more acceptable when the context provides a salient discourse goal supporting the MS interpretation. If MS is marked, then its acceptability should be contingent on the availability of this special contextual licensing. Therefore, removing the context should yield decreased acceptability of MS.

The second contextual factor involves how the the domain of reference of the *wh*-word is fixed. This factor derives from the observation that MS/MA variation is conditioned by the kind of *wh*-question , by the *wh*-word. Ginzburg (1995); Asher and Lascarides (1998) note that *who*-questions are MA-biased, while non-*who*-questions are not (and possibly MS-biased):

(6) a. Who came to the party? MA
b. Where can I find coffee? MA or MS
c. How do I get to the buried treasure? MS

Incidentally, support for an MA semantics comes mostly from *who*-questions (Karttunen, 1977; Groenendijk & Stokhof, 1982, 1984), while support for MS semantics comes mostly from non-*who*-questions (for example, *how*-questions factor importantly in Hintikka, 1974; Ginzburg, 1995; Asher & Lascarides, 1998)[mm: right date for hintikka?] (a point first made by Asher & Lascarides, 1998).

There are several other well-known ways in which aspects of fixing the *wh*-domain influences the meaning of *wh*-questions . Ginzburg notes that *wh*-words can differ in the

default granularity of their referential domains. Consider: A similar example, showing granularity effects, is given by Ginzburg:

(7) Context 1: Jane gets off the airport in Helsinki.
   a. Flight Attendant: Do you know where you are?
   b. Jane: Helsinki.
   c. Flight Attendant: Jane knows where she is.

(8) Context 2: Jane steps out of the taxi from the airport to her hotel in Helsinki.
   a. Taxi Driver: Do you kow where you are?
   b. Jane: Helsinki.
   c. Taxi Driver: Jane doesn't know where she is.

In the two examples, the granularity of reference for *knowing where one is* is fixed by Jane's answer "Helsinki." The reason that the taxis driver denies Jane's knowledge in the second case is because the level of granularity is discordant with what is necessary given the goals of the context. In other words, she is required to have a more *fine-grained* knowledge-*where* in order to navigate inside the city.

Filling out the logic, *who*-questions default to an individual-level granularity, while for *where*-, *how*-, *why*-questions there is no clear default. For those latter kind, context provides information that specifies the granularity of the domain. Asher & Lascarides take these considerations further by arguing that it is not cognitively reasonable to require exhaustivity in the case of those later *wh*-questions because usually the domains cannot be *a priori* constrained enough to a reasonable size.[2]

Similarly, classic issues of *de re*/*de dicto* ambiguities arise in identity questions as shown below ((Boër & Lycan, 1975)[mm: boer and lycan 1985, Quine?](Heim, 1994; Groenendijk & Stokhof, 1984)).

(9) Who is Cassius Clay?
   a. Mohammad Ali. *de dicto*
   b. That guy [pointing at Cassius Clay]. *de re*

(10) Scully knows who Cassius Clay is.

If Scully is taking an exam, then (9a) is a better answer than (9b); Yet if Scully wants to interview Cassius Clay at a boxing match, then (9b) is a better answer than (9a). The truth of (10) will depend on the match between context and reading.

Finally, implicit domain restriction is a general phenomenon by which speakers and hearers implicit restrict the domain of reference of a quantifier ((von Fintel, 1994)). For instance, if Mulder tells Scully of a party she missed, that *Everyone was at the party*, Mulder doesn't literally mean that everyone in the world was at the party. Rather, he meant that every *relevant* person—likely every mutual acquaintance–attended the party. The referential domain of *everyone* is implicitly restricted to a subset domain, *every relevant person*. Some have argued that these facts about MS are due to implicit domain restriction on the *wh*-word. (see (George, 2011), section 6.2.3, pp.211-214; [mm: fox 2018]).

[2]According to Dayal (p.c.), this kind of problem is solved by appeal to intensions in a possible world semantic framework because the domain need not be specified extensionally.

Generally, semantic theories do not attribute MS/MA variation to *wh*-word; *wh*-words are treated consistently except for the particular differences due to the *wh*-domain's range (i.e., roughly *who* ranges over people or individuals, *where* over places, *when* over times, *why* over reasons, *how* over ways). However, [mm: Aloni 2001, 2005] provides a formal account for the pragmatics of many of these issues in the method of identification of a *wh*-word . Her "conceptual covers" approach argued for an index on the *wh*-word that is fixed contextually relative the mental state of the speaker, following arguments by [mm: boer and lycan] and Ginzburg. Independently, Lahiri (2002) also for context-sensitive variables which allows (roughly) a convert quantifier to vary between existential and universal force, deriving MS and MA respectively.

Finally, there has been some discussion about the role that d-linking and number marking may play (Pesetsky, 1987; Dayal, 1996; Comorovski, 1996; Xiang, 2016)[mm: comorovski 2006?, sirivastav 1991,]. Dayal (2016) aruges convincingly that these facts can be overidden by context. Xiang & Cremers (2017) tested the availability of MS in plural-marked *wh*-questions like *Mary remembers which children can lead the dance* as compared to *Mary remember who can lead the dance*. In Experiment 1, they found no main effect of *wh*-word or interaction with modality, while in Experiment 2 they only found a significant interaction with modality.

Recent experimental work (Moyer, 2020; Moyer & Syrett, 2019) has confirmed that the *wh*-word does indeed modulate interpretation as predited by Ginzburg and Asher & Lascarides: MS contexts were less acceptable for *who*-questions than for *where*-questions. At the same time, there was variation based on particular items suggesting that this variation in *wh*-word is really tracking contextual modulation in how the *wh*-word domain is fixed.

Thus, if non-*who* questions require more contextual specification than *who*-questions,

Prediction 5: *wh*-word asymmetry.
On average, *who*-questions should exhibit an MA bias, while non-*who* questions should on average be more acceptable with MS interpretations.

2.3 Summary of predictions

Here we summarize the observations and predictions.

	Prediction
Overall	On average, *wh*-questions will be interpreted MA more than MS.
Modality: Presence **Modality: Absence**	On average, modal questions will be interpreted MS. On average, non-modal questions will be interpreted MA.
Matrix Verb	On average, *know-wh* will be interpreted MA.
Wh-Word	On average, *who*-questions will be interpreted MA, non-*who*-questions will be interpreted MS.
Discourse Goals	MS/MA interpretations require crucial information provided by the discourse context.

Table 1: Summary of Main factors of influence on MS/MA interpretation, and the directionality of their predicted influence.

3 The database

We used TGrep2 (Rohde, 2005) and the TGrep2 Database Tools (Degen & Jaeger, 2011) to extract 10,192 occurrences of utterances containing a *wh*-phrase from the Switchboard corpus of spoken American English (Godfrey, Hilliman, & McDaniel, 1992), a corpus of telephone conversations about assigned topics between strangers. Each utterance was annotated automatically for features of interest, including presence/absence of modality, *wh*-word , and syntactic structure (e.g., embedded vs. root question).

Since our goal was to investigate the interpretation of questions for which the issue of whether they receive an MS or MA interpretation can arise in the first place, we excluded all instances of degree questions (e.g., *How old are they?*), questions with complex *wh*-phrases (e.g., *Which group do you work in?*), and identity questions (e.g., *Who is their quarterback?*). These questions have only one interpretation (in which MS and MA converge). [jd: for each of these, say how many cases there were]

The remaining [jd: XXX] cases included [jd: XXX] root and [jd: XXX] embedded questions. To avoid data sparsity issues we focused on just the *wh*-questions headed by *who*, *what*, *where*, *when*, *how*, and *why*, leaving [jd: XXX] cases of root and [jd: XXX] cases of embedded questions. Table 2 presents the joint distribution of *wh*-words and the presence of modal auxiliary verbs in the database [jd: add the embedded cases. this section should report the whole database]. *What*-questions comprise 58.8% of this constrained set, followed by *how*-questions at 18.5%. [jd: update to also reflect embedded questions]

Experiments 1a and 2a tested interpretation of root questions; Experiments 1b and 2b tested interpretation of embedded questions.

[jd: the database section could also contain some examples of root and embedded questions to give people an intuition for what to expect – these can be some of the examples you discuss later on, ideally typical examples of +/- modal, +/- embedded, and different wh cases]

Table 2: Distribution of *wh*-words and modality in Switchboard root questions. Percentage of total (995).

.

wh-word	Root questions		Embedded questions	
	Modal present	Modal absent	Modal present	Modal absent
What	6.7%	52.1%	8%	38.2%
How	2.8%	15.7%	13%	14.9%
Where	0.4%	9.3%	1.7%	6.9%
Why	1.3%	4.7%	1.8%	7%
Who	0.8%	4.7%	0.5%	5.5%
When	0.1%	1.4%	0.48%	1.9%

4 Experiment 1: questions in context

In this set of experiments,[3] participants judged the exhaustivity of root (Exp. 1a) and embedded (Exp. 1b) questions presented in context using a paraphrase rating task (Degen, 2015).

4.1 Experiment 1a: root question interpretation

4.1.1 Method

Participants On Prolific, we recruited 660 speakers who were paid an average rate of $14/hr for their work. Eligible participants had to be born and currently reside in the US, as well as speak English as their first language. We excluded [mm: XX] participants who reported a native language other than English. We additionally removed 35 participants for failing 2 out of 6 control trials, and 4 participants whose data were not recorded due to browser error.

Procedure and materials On each trial, participants saw a question in the context of its 10 preceding lines of dialogue, and rated the speaker's likely intended meanings by adjusting a continuous slider for each of three paraphrases of the question (see Fig. 1 for an example). Paraphrases were constructed to reflect MS and MA readings: for instance, for the question "so, where have you skied?", the (exhaustive) MA reading and (non-exhaustive) MS reading were paraphrased as "What is every place you have skied?" and "What is a place you have skied?", respectively. We also included an additional paraphrase intended to capture cases where the MS and MA readings converge due to the uniqueness presupposition introduced by the definite determiner (Dayal, 1996)[mm:

[3] Procedure, materials, analyses and exclusions were pre-registered at `https://bit.ly/3tp1FC1`. Experimental materials and analysis scripts are available at [jd: insert github repo. you'll have to anonymize these links for review though]

Speaker #**2**: pretty good.
Speaker #**1**: i do like to ski.
Speaker #**2**: pretty, pretty down there. huh?
Speaker #**1**: yeah, i , i said i do like to ski.
Speaker #**2**: so, where have you skied?

Based on the sentence in red, how likely do you think it is that the speaker wanted to know about each of the following?

Figure 1: Example trial in Exp. 1a. [jd: would be neat if this screen shot had the sliders placed at the actual mean values for that item]

higginbotham and may 1981]. For instance, the single answer to "What is the place in which you have skied?" is both MS (because it is a single answer) and MA (because it answers the question exhaustively). A fourth option ("something else") was presented in case none of the paraphrases was appropriate.

Paraphrases were constructed based on the *wh*-word domain, and can be seen in the table below.

what	what was {a / the / every} place
how	what was {a / the / every} way
why	what was {a / the / every} reason
where	what was {a / the / every} place
who	who was {a / the / every} person
when	what was {a / the / every} time

Table 3: Paraphrases for each *wh*-word matched the domain and animacy restrictions imposed by the *wh*-word.

The 995 root questions in the database were randomly divided into 32 lists of 30 questions, and 1 list of 35 questions. Within each list, the distribution of *wh*-words and modals was kept roughly proportional to that of the overall set of root questions, so that each participant rated 30 questions which were representative of the overall question distribution in the linguistic features of interest.

In addition, each participant rated 6 control items. Similar to test items, control items were dialogues where an interlocutor asked a question. Rather than a *wh*-question ,

a polar question was instead asked. These polar questions were created to control for the three paraphrases by containing either an indefinite, definite, or universally quantified noun phrase, similar to the three paraphrases. As such, they are intended to unambiguously convey the meaning encoded in a paraphrase. For example, an *a*-control question like *Can you grab a tissue?* should give rise to highest ratings for the *a*-paraphrase because the speaker explicitly states that they want *a tissue*.

Participants first completed four training trials with example dialogues: on two of these, the *a*/*the* paraphrases were best [jd: why both?], and on the other two the *every* paraphrase was best. We included one modal and one non-modal question for each. Further, on each training trial, participants were instructed to interpret the ellipsis in each paraphrase relative to the content in the target question. For instance, this yields "What is every place you have skied?" from the displayed "What is every place. . . ?" in Fig. 1.

4.1.2 Exclusions and preprocessing

Questions with higher mean ratings for *something else* than any other option were removed (15%). These tended to be rhetorical questions (e.g., *Who knows?*, *Who has the time?*, *What are we becoming?*), whose interpretation is orthogonal to the question of whether *wh*-questions are interpreted exhaustively. After exclusions, [jd: XXX] cases remained for analysis. For each item and participant, ratings were normalized such that the three remaining slider values summed to 1.

4.1.3 Results

Because this is a novel task for testing *wh*-question interpretation, we first report a qualitative analysis of the data to assess whether the ratings given for particular items accord with intuitions. We then report the main analysis of interest, which assesses Predictions 1-3 and 5 (all predictions except Goal Sensitivity).

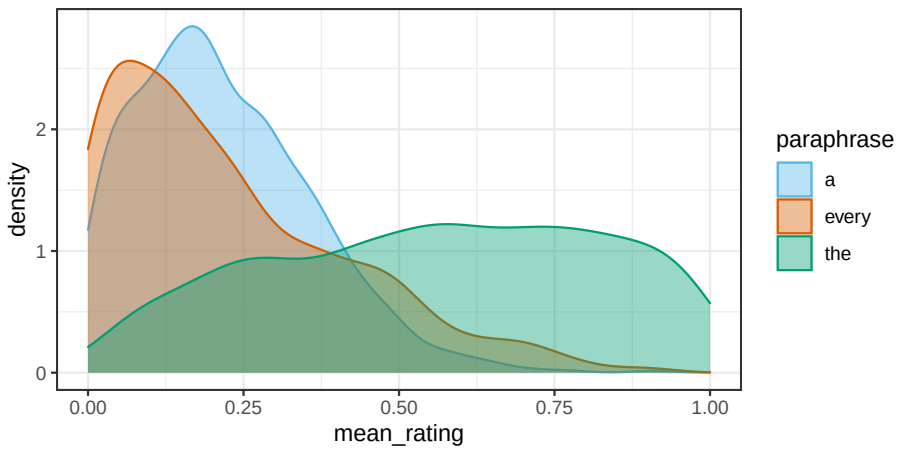

Figure 2: Mean ratings by item in Experiment 1a.

	Question	Mean	SD
(a)	*Where do you live?*	1	0
(b)	*Where do you work?*	.99	.03
(c)	*What do you drive now?*	.99	.0)
(d)	*How do you spell that?*	.99	.0
(e)	*When did you first take your first piano lesson?*	.92	.24
(f)	*Why are you cutting off the phone?*	.81	.36
(g)	*What does it have in it?*	.86	.31
(h)	*Where have you skied?*	.73	.37
(i)	*What is a good brand, a inexpensive?*	.63	.39
(j)	*What have you seen lately?*	.66	.3
(k)	*How can I tell you?*	.66	.42
(l)	*What else can we talk about?*	.99	.21

Table 4: A selection of questions rated highest for each rating. The first group contains questions rated highest on the *the*-paraphrase, the second for the *every*-paraphrase, and the third for the *a*-paraphrase.

Qualitative analysis For each paraphrase, we assesed the questions rated highest for that paraphrase. We refer to 4. Items (a) - (f) received high *the*-paraphrase ratings (at or near 1). These are all questions that are indeed highly likely to have exactly one answer. Items (g)-(h) are questions that received high *every* ratings. The first occurred in a context about cooking a casserole; the second in a conversation about the hearer's love for skiing. The exhaustive interpretation—wanting to know all the ingredients in the casserole, wanting to know all the places the hearer has skied—is sensible in both cases.

Questions that received high *a* ratings often involved recommendations, e.g., (i), which occurred in a discussion about computers where the hearer was an expert. Many questions involved discussions about books or movies, e.g., (j). Other interesting cases included (k), where the speaker struggled to articulate (tell) why they like a certain movie, and (l) where the speaker is struggling to find a conversation topic. In both cases, there are presumably multiple answers (ways to tell, things to talk about), but a single one is sufficient to achieve the speaker's goal.

Overall, the qualitative assessment of individual items suggests that participants understood the task and that the paraphrase ratings are interpretable as proxy judgments for *wh*-question exhaustivity.

Quantitative analysis Analyses were conducted on the subset of the *a*/*every* paraphrase data, to assess overall question interpretation bias and the effect of modality and *wh*-word on question interpretation. To this end, we conducted a mixed effects linear regression predicting ratings from fixed effects of WH-WORD (reference level: *when*), centered measures of whether a modal auxiliary verb was present (MODALPRESENT, before centering: 'not present'=0, 'present'=1), and PARAPHRASE (before centering: 0 = *every*, 1 = *a*), all 2-way interactions, and the 3-way interaction. The model included the maximal

random effects structure justified by the design: random by-item and by-subject intercepts, as well as by-item and by-subject slopes for PARAPHRASE, and by-subject slopes for WH and MODALPRESENT.

We observed significant 3-way interactions. However, interpreting the interaction terms in this full model is very complex. We thus take the significant three-way interactions as evidence that effects varied by *wh*-word and report the outcome of separate *wh*-word -specific models on each *wh*-word subset of the data: each model included fixed effects of PARAPHRASE, MODALPRESENT, and their interaction, coded as in the full model.

Prediction 1: Is there an overall MA bias? Fig. 4 [jd: why is this figure made reference to first, if you show us three other plots before it? i'm generally confused by the plot choices, let's discuss] shows mean ratings for each paraphrase. Rather than a preference for MA over MS readings, there was a clear preference for the *the*-paraphrase. There was only a significant MA bias [jd: first say which coefficients you interpret as evidence for MA/MS bias – perhaps you can merge this in with where you say that the data analysis will only be conducted on the subset of a/every paraphrase ratings] for *what* questions (β=-0.04, SE=0.01, t=-3.13, p <0.002). In contrast, for *how* (β=0.06, SE=0.02, t=4.02, p <0.0001), *why* (β=0.06, SE=0.02, t=2.71, p <0.009) and *when*-questions (β=0.15, SE=0.03, t=5.47, p <0.0001) there were significant MS biases. Finally, for *where* and *who* there were no significant biases in either direction.

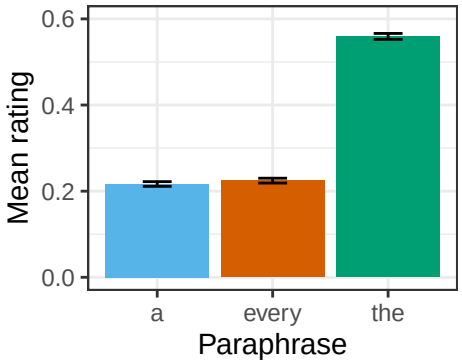

Figure 3: Mean ratings by paraphrase in Experiment 1a. Here and below, error bars indicate 95% bootstrapped confidence intervals. [jd: include the "something else" ratings in this plot for completeness. add a visualization of variance (eg, plot means as dots with error bars overlaid on violin plots)]

4.1.4 Predictions 2 and 5: Do modality and *wh*-word modulate question interpretation?

[jd: re-organize this part to fit with the predictions listed in intro?] Simply, yes. Fig. 4 presents mean ratings as a function of paraphrase and the presence of a modal, separately for each *wh*-word . We focus the following discussion on the coefficients and p-values in the Table 5.

Table 5: Coefficient table (predicted β coefficient, standard error SE, t value, and p value) for *wh*-word -specific models in Exp. 1a. [jd: unify reporting so each beta, se, and t value shows exactly 2 decimal point (ie, insert missing zeroes. also, get rid of vertical lines)]

Wh-Word		β	SE	t	p
WHAT	Intercept	.25	.006	38.74	$<$.0001
	ModalPresent	.09	.02	4.76	$<$.0001
	Paraphrase	-.04	.01	-3.13	$<$.002
	ModalPresent:Paraphrase	.09	.03	2.8	$<$.006
HOW	Intercept	.20	.01	24.63	$<$.0001
	ModalPresent	.08	.02	3.62	$<$ .0005
	Paraphrase	.06	.02	4.02	$<$.0001
	ModalPresent:Paraphrase	.15	.04	3.89	$<$.0002
WHERE	Intercept	.13	.01	9.35	$<$.0001
	ModalPresent	-.03	.08	-.37	.72
	Paraphrase	-.01	.02	-.43	.6
	ModalPresent:Paraphrase	.05	.09	.52	.6
WHY	Intercept	.21	.01	21.57	$<$.0001
	ModalPresent	.02	.02	.87	.34
	Paraphrase	.06	.02	2.71	$<$.009
	ModalPresent:Paraphrase	.09	.05	1.82	.07
WHO	Intercept	.2	.03	7.6	$<$.0001
	ModalPresent	.1	.06	1.6	.12
	Paraphrase	.04	.05	.79	.43
	ModalPresent:Paraphrase	.2	.11	1.71	.09
WHEN	Intercept	.12	.02	6.41	$<$.0001
	ModalPresent	.28	.08	3.5	$<$.004
	Paraphrase	.15	.03	5.45	$<$.0001
	ModalPresent:Paraphrase	.54	.12	4.71	$<$.0002

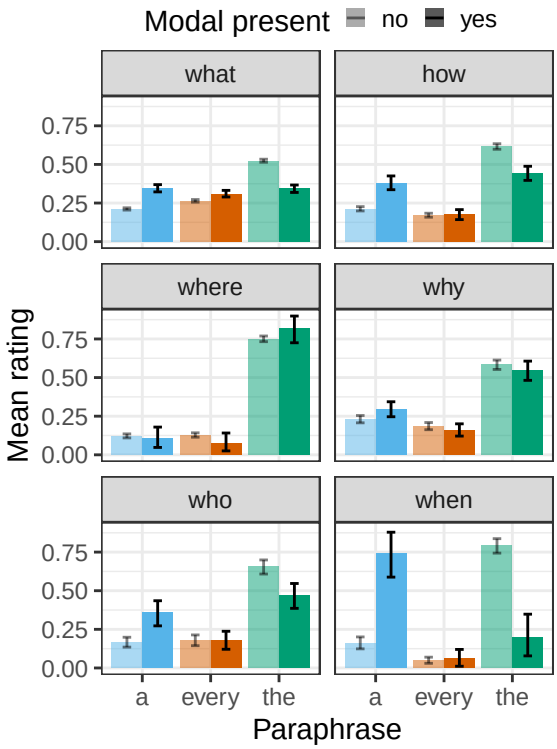

Figure 4: Mean ratings by paraphrase, *wh*-word and modality in Experiment 1a. Error bars indicate 95% bootstrapped confidence intervals.

For *what* questions, the overall MA bias is inverted to an MS bias when a modal is present. [jd: provide two examples that show the contrast?]

How, *why*, and *when* pattern together in showing an overall MS bias, confirming previously made predictions (Ginzburg, 1995; Asher & Lascarides, 1998). *Who* and *where* show no overall bias, but *who* shows a preference for MS while *where* for MA, in contrast to those predictions. [jd: this last sentence is unparseable]

Thus, the prediction that *who* is biased for MA is not confirmed. Only *what* displayed an MA bias. Qualitative inspection of those questions which received the highest *every* ratings revealed that some of these were questions with a plural-marked complex *wh*-phrase (e.g., *What cities are they looking at*, .92, .15) that slipped through our initial filters. These were intended to be excluded precisely because they are expected to not be ambiguous between MS and MA. [jd: again, this makes it sound like we're just trying to ecxlude cases that don't work in our favor. explain why that's not the case] However, future work should explicitly include and test complex *wh*-questions. If they really are unambiguous, then plural-marked complex questions should show a clear preference for *every*, and singular-marked ones for *a* (or *the*).

The presence of a modal yielded higher MS ratings for *what*, *how*, and *when* questions, while this increase was only marginally significant for *why* and *who* questions, and did not reach significance for *where* questions. [jd: i'm confused – why are the interactions not reported, even though they're doing the main work?]

Overall, these results confirm the observation that modal auxiliaries facilitate MS read-

ings.

4.2 Experiment 1b: embedded question interpretation

Exp. 1b tested exhaustivity of questions embedded under a variety of predicates, some of which have been claimed to semantically condition exhaustivity [jd: cite].

[jd: edit this whole section so the methods and results sections mirror the re-organization of exp 1a]

4.2.1 Method

Participants On Prolific, we recruited [mm: 1073–is this right?] speakers who were paid about $14/hr for their work. Eligible participants had to be born and currently reside in the US, as well as speak English as their first language. We included an addition question at the end of the study about native languages, and excluded 25 participants who reported native languages other than English. We additionally removed 37 participants for failing 2 out of 6 control trials.

Procedure and materials The procedure and materials were nearly identical to that of Experiment 1a, except in two respects. Figure Fig. 5.

First, the materials were embedded questions from the Switchboard corpus rather than root question. The 1075 embedded question database was divided into 35 lists of 30 questions, and 1 list of 35 questions. The distribution of *wh*-words and modals was kept roughly proportional to the overall distribution.

Second, and in virtue of the first difference, the task itself changed in two ways. First, participants responded to the question, *Based on the question in red, how likely do you think it is the that speaker means each of the following?*. Second, the paraphrases included ellipses before the *wh*-word as well, to indicate that the paraphrase involved the lingusitic content occuring prior to the *wh*-word .

4.2.2 Exclusions and preprocessing

[mm: what is the equivalent of a rhetorical question in embedded cases?] Questions that received higher ratings for *something else* than any other option were removed (15.5%). These are questions like “but they do read. uh, where a lot of people don’t have any interest in it at all,” “they’re real liberal now and to where probably fifty or a hundred years ago, um, the democrat party being liberal like they are?,” “but i like that where they run tense,” whose interpretation is orthogonal to the question of whether *wh*-questions are interpreted exhaustively. [jd: um, it’s not just an issue of the question of exhaustivity being orthogonal, it’s also that these clearly include cases that shouldn’t have been included in the first place. also, these aren’t readable if they just occur one after the other. put in between quotes instead of italicizing and be explicit about the issue with the sentences.] After exclusion, ratings were normalized such that for each participant and item, the three remaining slider values summed to 1.

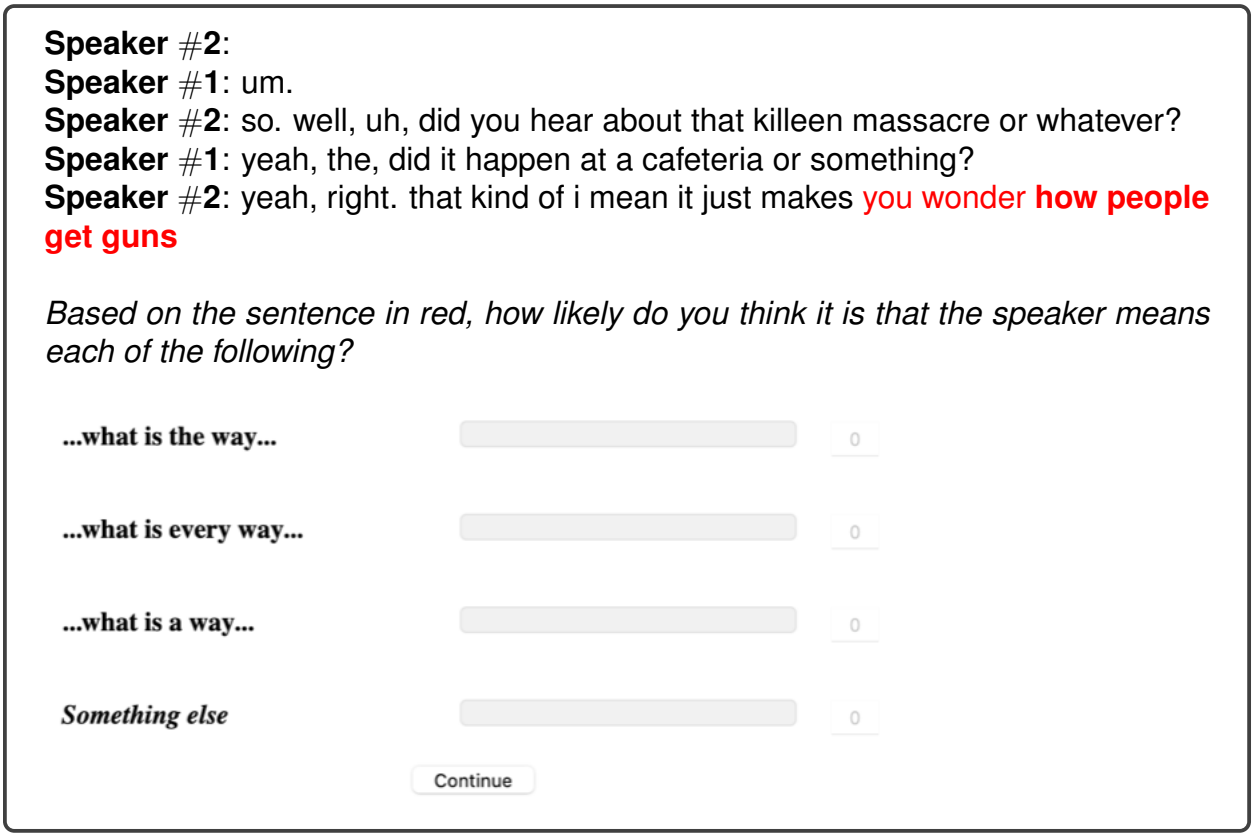

Figure 5: Example trial in Exp. 1b.

4.2.3 Qualitative analysis

Because this is a novel task for testing *wh*-question interpretation, we begin by qualitatively assessing whether the ratings given for particular items accord with intuitions about the best paraphrase.

Questions like "I know who you're talking about," (.94,.13), "Do you know who the guy was that was playing the wagon driver?" (.95,.14), "My mother taught me how to make it," (.93,.2), and "I forget how she got it," (.86,.14) all received high mean ratings for *the*-paraphrase. For these questions, it is again possible but unlikely that there is more than one answer.

Questions that received a high rating for the *every*-paraphrase included "So He knew who worked there," (.79,.35), "I have got to be so much more careful with what I do," (.77,.35), "Nobody can predict what's going to happen in twenty years," (.69,.38). In the first case, the preceeding discourse context included an explicit domain restriction (*there's only a few people that worked there*), making an MA reading possible. The second question occurs in a context about an injury, where the necessissity for caution requires MA. Finally, the third question occurs in the context of planning for possible events so an *every* reading is salient.

Questions that received a high rating for the *a*-paraphrase included "I'm not real sure why anybody would need a full automatic weapon," (.61,.4), "I don't know how to make it better for them," (.57,.4), and "I don't know when I'm going to get them all," (.49,.4). In all these three cases, an MS interpretation is sufficient to achieve the speaker's goal.

Overall, the qualitative assessment of individual items suggests that participants understood the task and that the ratings are interpretable.

4.2.4 Data analysis

Analyses were conducted to assess overall question interpretation bias and the effect of modality and *wh*-word on question interpretation. To this end, we conducted a mixed effects linear regression predicting rating from fixed effects of PARAPHRASE (reference level: *every*), WH-WORD (reference level: *when*), a dummy-coded and mean-centered measure of whether a modal auxiliary verb was present (MODALPRESENT), all 2-way interactions between fixed effects, and the 3-way interaction. We included the maximal random effects structure justified by the design: random by-item and by-subject intercepts, as well as by-item and by-subject slopes for PARAPHRASE, and by-subject slopes for WH and MODALPRESENT.

We observed significant 3-way interactions. However, interpreting the interaction terms in this full model is very complex because two of our predictors include > 3 levels. We thus take the significant three-way interactions as evidence that effects varied by *wh*-word and report the outcome of separate specific models on each *wh*-word subset of the data: each model included fixed effects of PARAPHRASE, MODALPRESENT, and their interaction, coded as in the full model.

An exhaustive MA bias is evidenced as a significantly negative coefficient of the 'a vs. every' PARAPHRASE contrast; a non-exhaustive MS bias as a significantly positive coefficient. An introduction or strengthening of an MS bias in the presence of a modal is

evidenced in a significantly positive interaction of MODALPRESENT with the 'a vs. every' PARAPHRASE contrast. The results of each model are shown in Table 6, with the two relevant contrasts highlighted in gray.

4.2.5 Results

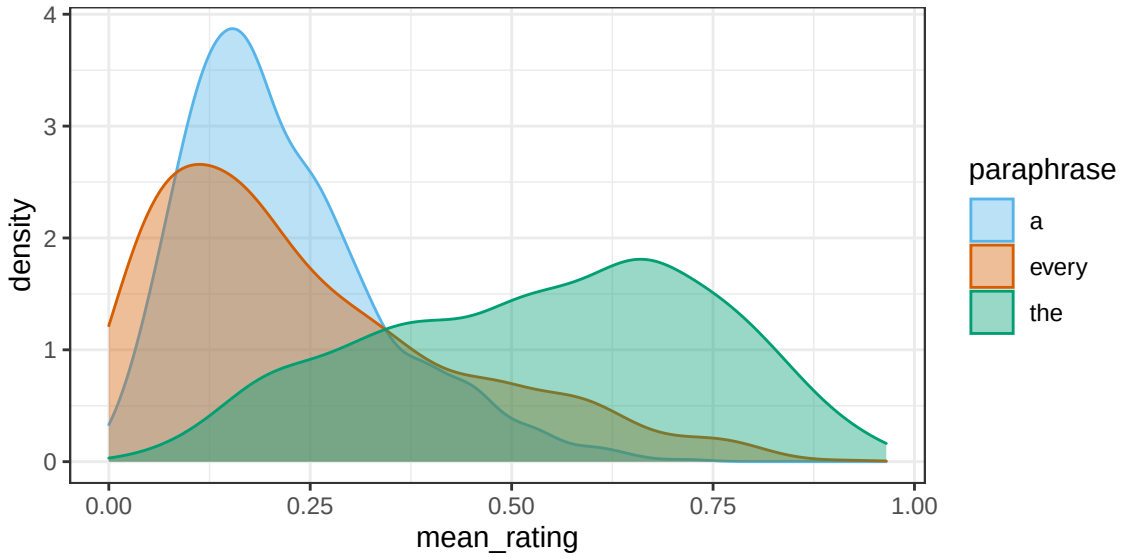

Figure 6: Mean ratings by item for *a*- and *every*-paraphrases in Experiment 1b.

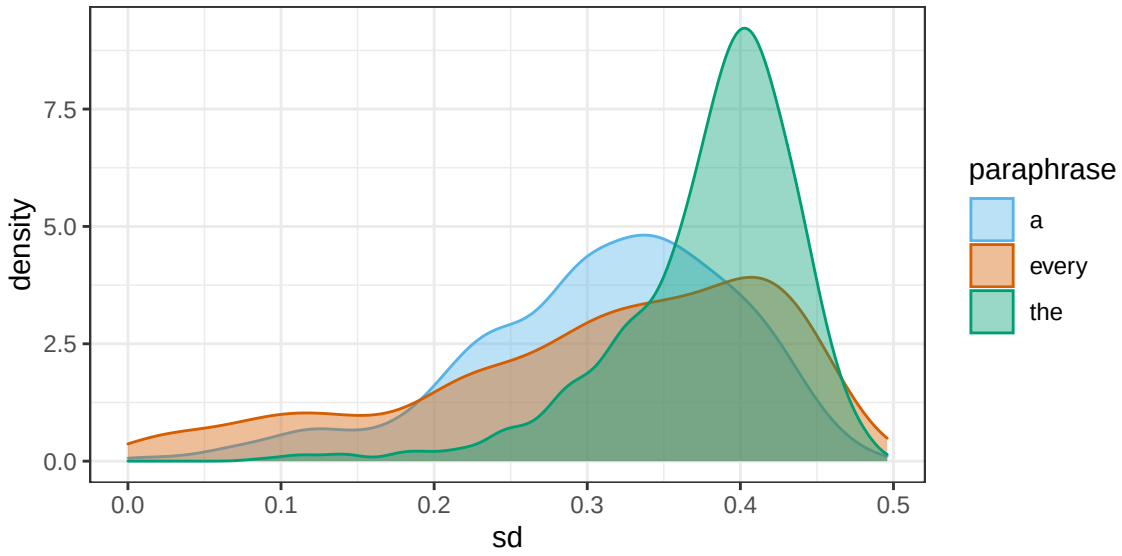

Figure 7: SD by item for *a*- and *every*-paraphrases in Experiment 1b.

Table 6: Coefficient table (predicted β coefficient, standard error SE, t value, and p value) for *wh*-word -specific models in Experiment 1b. Both predictors are dummy-coded and centered (0 is no Modal Present and *every*-paraphrase, 1 for Modal is present and *a*-paraphrase.

Wh-Word		β	SE	t	p
WHAT	Intercept	.26	.01	22.85	$<$.0001
	ModalPresent	.07	.01	5.83	$<$.0001
	Paraphrase	-.12	.02	-6.51	$<$.0001
	ModalPresent:Paraphrase	.09	.03	2.77	$<$.006
HOW	Intercept	.22	.01	19.21	$<$.0001
	ModalPresent	.03	.01	2.37	$<$.02
	Paraphrase	.06	.01	3.99	$<$.0002
	ModalPresent:Paraphrase	0.09	.02	4.12	$<$.0001
WHERE	Intercept	.21	.02	13.69	$<$.0001
	ModalPresent	0.03	0.03	1.09	.28
	Paraphrase	.01	.03	.22	.83
	ModalPresent:Paraphrase	.2	.08	2.42	$<$0.02
WHY	Intercept	.19	.01	27.9	$<$.0001
	ModalPresent	0.07	.02	4.06	$<$.0002
	Paraphrase	.05	.02	2.42	$<$.05
	ModalPresent:Paraphrase	.1	.04	2.51	$<$.02
WHO	Intercept	.25	.02	12.08	$<$.0001
	ModalPresent	.08	.07	1.15	.26
	Paraphrase	-.05	.06	-.82	.42
	ModalPresent:Paraphrase	-.01	.14	-.04	.99
WHEN	Intercept	.23	.04	6.37	$<$.0001
	ModalPresent	.03	.06	.51	.62
	Paraphrase	.12	.06	2.05	.07
	ModalPresent:Paraphrase	.14	.14	1.05	.31

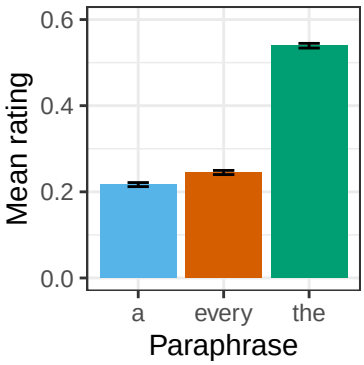

Figure 8: Mean ratings by paraphrase in Experiment 1b. Here and below, error bars indicate 95% bootstrapped confidence intervals.

Prediction 1: Is there an overall MA bias? Fig. 8 plots mean rating as a function of Paraphrase. We found evidence for a significant MA bias with *what*-questions (β=-.12, SE=.02, t=-6.51, p <0.0001), and a non-significant preference for MA with *who*-questions (β=-.05, SE=.06, t=-.82, p=.99). The remaining *wh*-questions revealed either significant bias (*how*: β=.06, SE=.01, t=3.99, p <0.0002; *why*: β=.05, SE=.02, t=2.42, p <0.05) or preference (*where*: β=.01, SE=.03, t=.22, p=.83; *when*: β=.12, SE=.06, t=2.05, p <0.07) for MS.

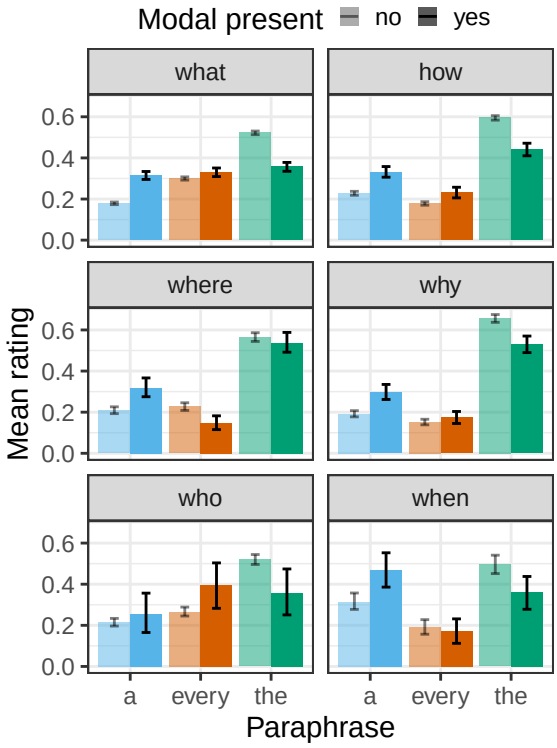

Figure 9: Mean ratings by paraphrase, *wh*-word and modality in Experiment 1b. Error bars indicate 95% bootstrapped confidence intervals.

Predictions 2,3 and 5: Is MS interpretation modulated by the presence of modality, by matrix verb, and by *wh*-word ? Yes. Fig. 9 plots mean rating as a function of Modal-Present, Wh-Word and Paraphrase. Interestingly, the presence of a modal significantly reversed the initial MA bias for *what*- questions (β=0.09, SE=0.03, t=2.77, p <0.006), but only attenuated the MA preference in *who*-questions. The remaining *wh*-questions patterned together, the presence of a modal strengthened the MS bias significantly (except for *when*-questions, where the effect did not reach significance).

Fig. 10 presents several matrix verbs that are of theoretical interest. Unfortunately, we encounter a sparse data problem for most of these verbs so we focus on *know*. Fig. 11 plots mean rating as a function of Paraphrase, Wh-Word and Modality. We found significant 2-way interactions between Wh-Word and Paraphrase in *know-wh*: a significant MA bias for *know-what* (β=-.09, SE=0.02, t=-4.75, p <0.0001) and a non-significant MA preference for *know-who* (β=-.1, SE=0.06, t=-1.77, p=0.09), while a signfifcant MS bias for *know-how* (β=.08, SE=.02, t=4.12, p <0.0001) and *know-when* (β=0.22, SE=0.09, t=2.41, p <0.05), and non-significant MS preference for *know-where*, *know-why*.

The addition of a modal either reversed an MA bias (*what*: β=0.12, SE=0.04, t=2.74, p <0.007) or strengthed the MS bias (*know*: β=0.01, SE=0.03, t=2.93, p <0.005, *where*: β=0.26, SE=0.01, t=2.58, p <0.03, *why*: β=0.19, SE=0.07, t=2.49, p <0.02). However, there was no effect of modal for *know-who*, *know-when* questions.

4.3 Discussion

Given our results, we conclude that indeed the MS reading of *wh*-questions is modulated by various linguistic factors. Figure Fig. 12 plots the distribution of mean ratings by item and by linguistic factors. As is clear from the plots, there is immese variation in interpretation.

The most robust finding is that the presence of a modal significantly increases ratings for MS than the absence. This effect held even though we grouped both necessity and possibility modals together. Looking more deeply at the distribution of readings with particular modals, [mm: add some discussion about modality here, possibility versus necessisity]

There were some slight differences between the two experiments. While there was no overall MA bias for root questions, there was more of a bias in embedded questions. In both cases, however, ratings were significantly modulated by the Wh-word and the presence of modality.

[mm: were the differences in the two experiments driven by know-wh? like, can we attribute the preference for MA in experiment 1b to know-wh? How would we test that? Can we test a model on the 1b data making a dummy coded variable for know-wh versus everything else?]

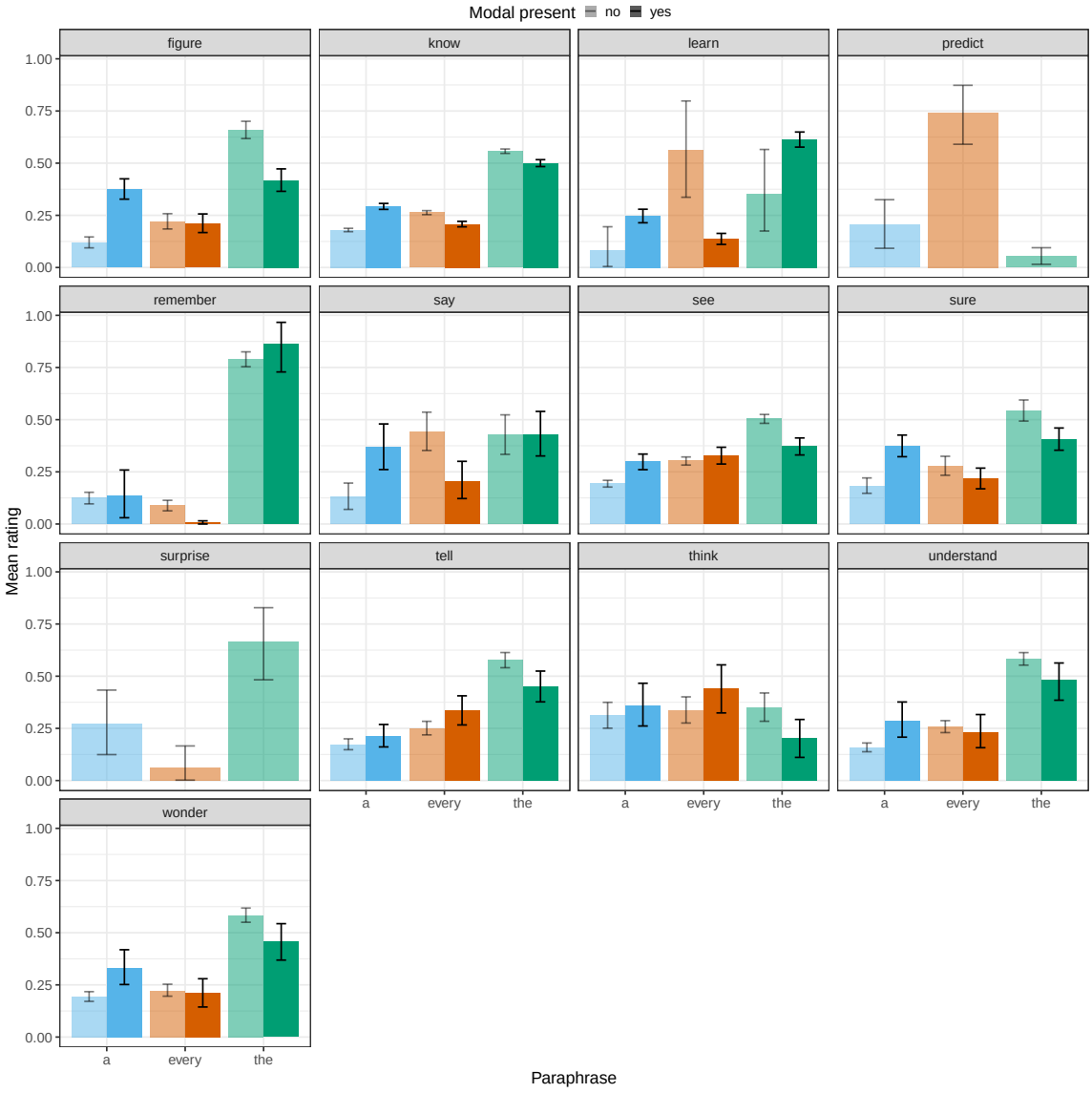

Figure 10: Mean ratings by paraphrase for several Matrix Verbs that have been of interest in the theoretical literature, in Experiment 1b. Error bars indicate 95% bootstrapped confidence intervals.

5 Experiments without context

We've shown that there is no MA bias, but that linguistic factors like both the Wh-word and Modality can influence the distribution of interpretiatons. It's possible that, by removing the contexts of utterance we understand the extent to which *wh*-question interpretation changes. In this second set of experiments, we conducted the same experiment with root and embedded questions above, however without the ten preceding lines of discourse.

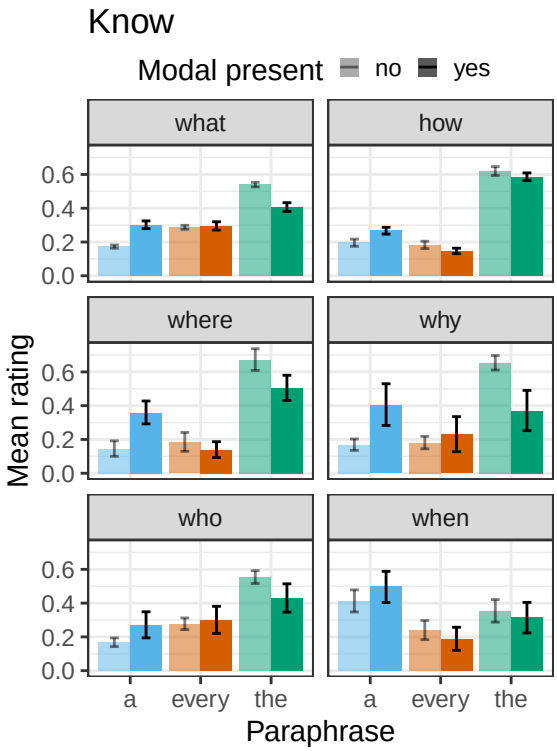

Figure 11: Mean ratings by paraphrase, *wh*-word and modality for questions embedded under *know* in Experiment 1b. Error bars indicate 95% bootstrapped confidence intervals.

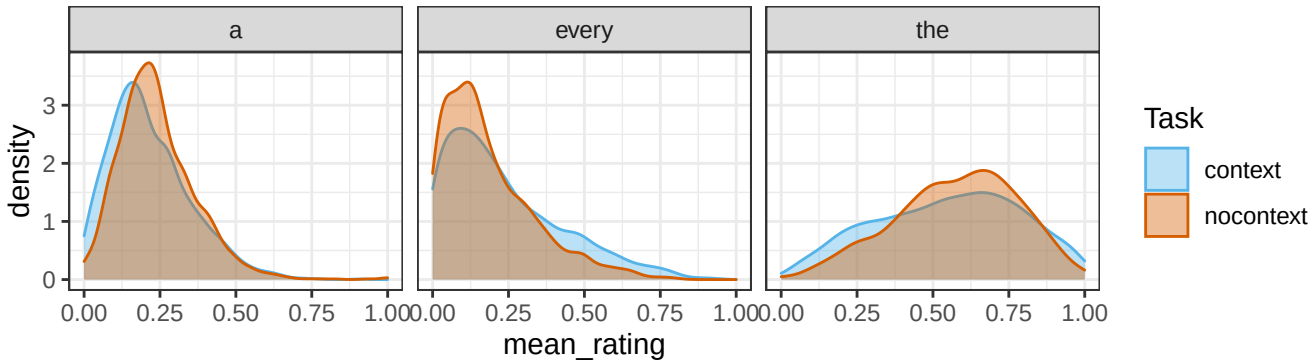

Figure 12: Mean ratings by item for *a*- and *every*-paraphrases, across Experiments 1a and 1b.

5.1 Predictions

Based on observations from the literature, we expect (1) questions to be overall biased for MA, (2) Modal questions to be biased for MS, and possibly non-modal questions for MA, (3) *who*-questions to be biased for MA, and others for MS. For experiment 1b with embedded questions, we predict that *know-wh* will be biased for MA, but also that there may be some interactions with *wh*-word .

Besides the predictions about overal and specific biases, we might expect differences between the Context and NoContext experiments if the discourse context provides crucial information relevant to determining the interpretation of root and embedded questions, which are underspecified for (non)-exhaustivity.

5.2 Experiment 2a: Root questions without context

5.2.1 Method

Participants 656 participants, 25 removed for non-native. 51 additional participants were removed for failing controls.

Procedure and materials The procedure and materials were nearly identical to that of Experiment 1a as presented in Fig. 1, except participants were shown the root questions without the ten preceeding lines of discourse.

5.2.2 Exclusions and preprocessing

Questions that received higher ratings for *something else* than any other option were removed (16.3%). After exclusion, ratings were normalized such that for each participant and item, the three remaining slider values summed to 1.

5.2.3 Data analysis

Analyses were conducted to assess overall question interpretation bias and the effect of modality and *wh*-word on question interpretation. To this end, we conducted a mixed effects linear regression predicting critical ratings from fixed effects of WH-WORD (reference level: *when*), dummy-coded and mean-centered measures of whether a modal auxiliary verb was present (MODALPRESENT, 0 = not present, 1 = present) and PARAPHRASE (0 = *every*, 1 = *a*), all 2-way interactions between fixed effects, and the 3-way interaction. We included the maximal random effects structure justified by the design: random by-item and by-subject intercepts, as well as by-item and by-subject slopes for PARAPHRASE, and by-subject slopes for WH and MODALPRESENT.

5.2.4 Results

Prediction 1: Is there an overall MA bias? Figure Fig. 14 plots the overal mean ratings for the three paraphrases. There were no significant biases for MA for any *wh*-question . *How*, *where*, *why*, *when* patterned together in showing significant MS biases, while *what* and *who* patterned together in having positive (but not significant) coefficiants, revealing a preference towards an MS bias.

Table 7: Coefficient table (predicted β coefficient, standard error SE, t value, and p value) for *wh*-word -specific models in Experiment 2a. Both predictors are dummy-coded and centered (0 is no Modal Present and *every*-paraphrase, 1 for Modal is present and *a*-paraphrase.

Wh-Word		β	SE	t	p
WHAT	Intercept	.23	.005	44.12	$<$.0001
	ModalPresent	.08	.01	5.43	$<$.0001
	Paraphrase	.02	.01	1.4	.16
	ModalPresent:Paraphrase	.02	.02	.81	.42
HOW	Intercept	.19	.01	29.07	$<$.0001
	ModalPresent	.05	.02	2.79	$<$.006
	Paraphrase	.09	.01	6.27	$<$.0001
	ModalPresent:Paraphrase	.07	.03	2.64	$<$.001
WHERE	Intercept	.13	.01	13.52	$<$.0001
	ModalPresent	-.01	.04	-.19	.85
	Paraphrase	.06	.01	5.04	$<$.0001
	ModalPresent:Paraphrase	-.1	.06	-1.64	.1
WHY	Intercept	.2	.01	24.32	$<$.0001
	ModalPresent	.03	.02	1.8	$<$.08
	Paraphrase	.08	.02	4.78	$<$.0001
	ModalPresent:Paraphrase	.06	.04	1.77	$<$.09
WHO	Intercept	.21	.02	10.46	$<$.0001
	ModalPresent	.03	.05	.62	.54
	Paraphrase	.03	.03	1.05	.3
	ModalPresent:Paraphrase	.2	.08	2.29	$<$.03
WHEN	Intercept	.11	.02	5.99	$<$.0001
	ModalPresent	.17	.06	2.71	$<$.02
	Paraphrase	.13	.03	3.73	$<$.003
	ModalPresent:Paraphrase	.04	.12	.29	.77

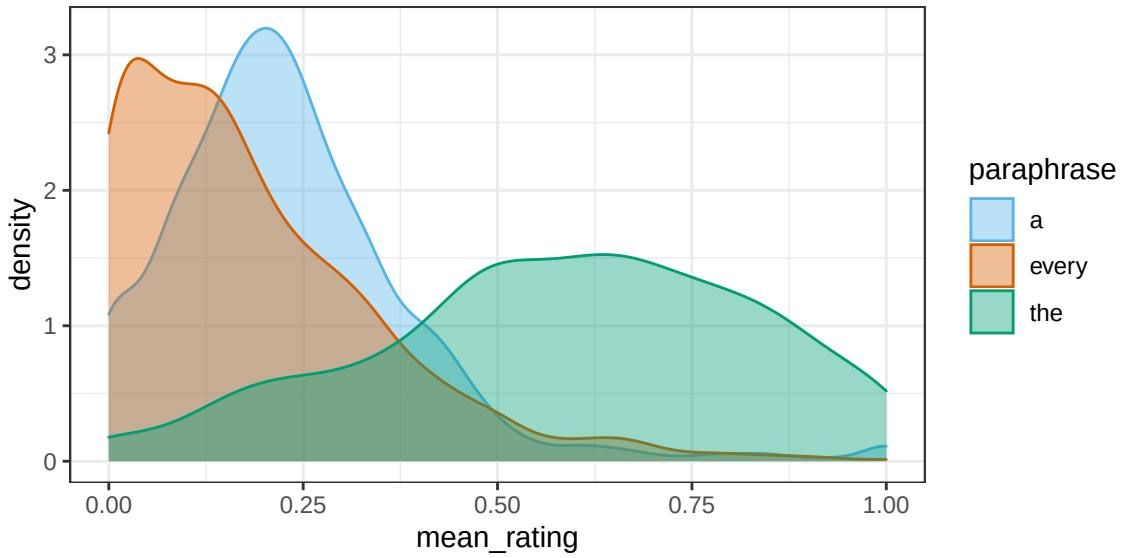

Figure 13: Mean ratings by item for *a*- and *every*-paraphrases in Experiment 2a.

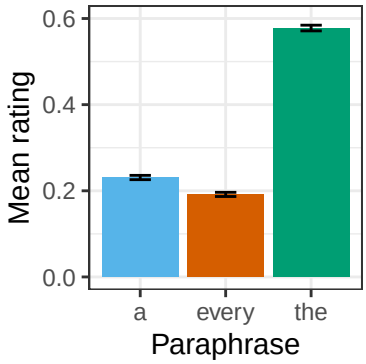

Figure 14: Mean ratings by paraphrase in Experiment 2a. Here and below, error bars indicate 95% bootstrapped confidence intervals.

Predictions 2 and 5: Is MS modulated by the presence of modality and *wh*-word ? Fig. 15 plots mean rating as a function of ModalPresent, Wh-Word and Paraphrase. As just discussed, there were some significant differences in initial biases for MS vs. MA based on the different *wh*-questions : *how*, *where*, *why*, *when* showed a significant MS bias, while *what* and *who* did not (though were positively skewed towards MS). This finding is different from the previous two studies, where at least *what* and *who* showed MA biases.

Given that there were no *wh*-questions biased for MA, the presence of a modal did not significantly reverse any biases, although for *when* and *where* the modal appeared to neutralize the MS bias. For *how*, *why* and *who* the interaction with modality was significant (and positively skewed towards MS), but not significant for *when*.

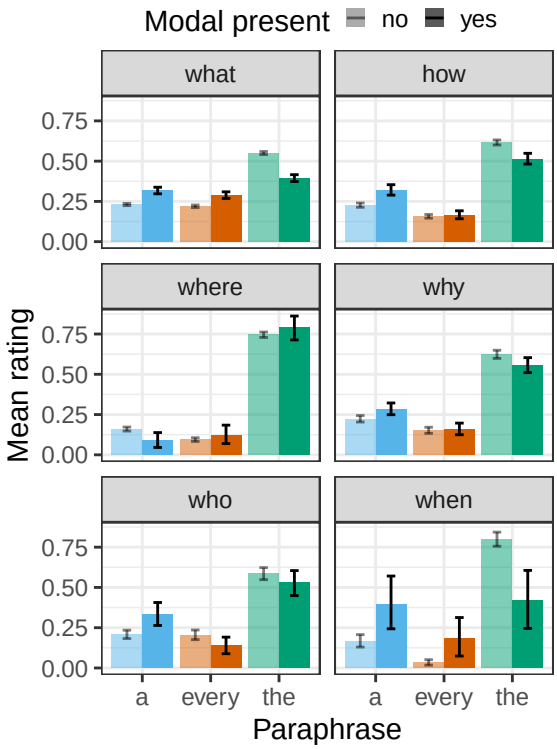

Figure 15: Mean ratings by paraphrase, *wh*-word and modality in Experiment 2a. Error bars indicate 95% bootstrapped confidence intervals.

5.3 Experiment 2b: Embedded questions without context

5.3.1 Method

Participants 717 participants, 28 removed for non-native, 43 removed for failing controls.

Procedure and materials The procedure and materials were nearly identical to that of Experiment 2b as presented in Fig. 5, except–as with Experiment 2a, participants were not shown the 10 preceeding lines of discourse.

5.3.2 Exclusions and preprocessing

Questions that received higher ratings for *something else* than any other option were removed (17.8%). After exclusion, ratings were normalized such that for each participant and item, the three remaining slider values summed to 1.

5.3.3 Data analysis

Analyses were conducted to assess overall question interpretation bias and the effect of modality and *wh*-word on question interpretation. To this end, we conducted a mixed ef-

fects linear regression predicting critical ratings from fixed effects of WH-WORD (reference level: *when*), dummy-coded and mean-centered measures of whether a modal auxiliary verb was present (MODALPRESENT, 0 = not present, 1 = present) and PARAPHRASE (0 = *every*, 1 = *a*), all 2-way interactions between fixed effects, and the 3-way interaction. We included the maximal random effects structure justified by the design: random by-item and by-subject intercepts, as well as by-item and by-subject slopes for PARAPHRASE, and by-subject slopes for WH and MODALPRESENT.

5.3.4 Results

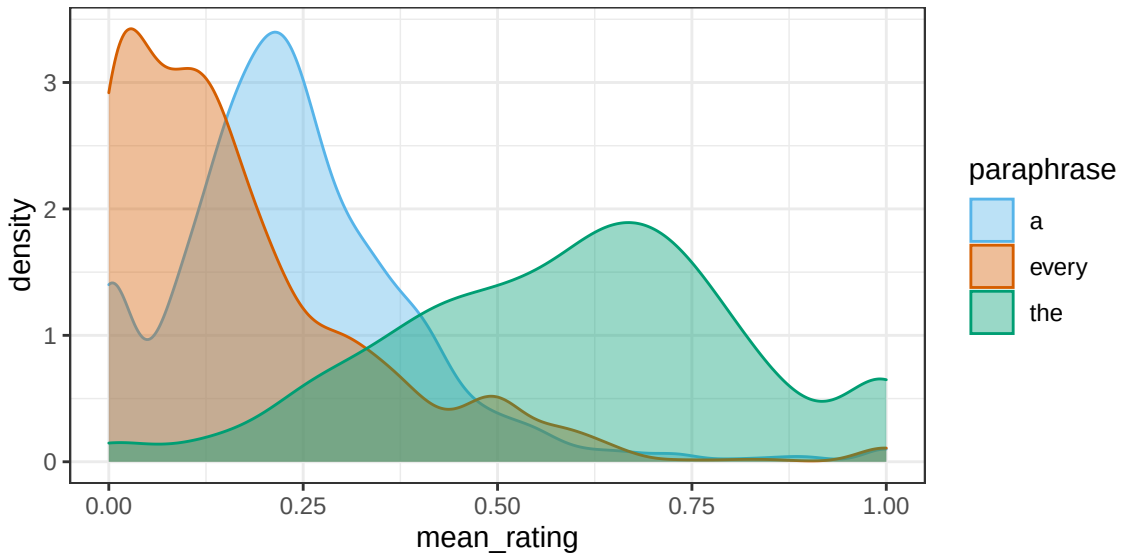

Figure 16: Mean ratings by item for *a*- and *every*-paraphrases in Experiment 2b.

Prediction 1: Is there an overall MA bias? Fig. 17 plots mean rating as a function of paraphrase. No *wh*-question revealed a significant initial bias for MA. In contrast, all *wh*-questions except *what*-questions revealed significant initial biases for MS paraphrases.

Predictions 2, 3, and 5: Is MS modulated by the presence of modality, matrix verb, and *wh*-word ? Fig. 18 plots mean rating as a function of ModalPresent, Wh-Word and Paraphrase. Unlike the previous three experiments, most questions patterned together in showing an MS bias, except *what*-questions which were consistently MA skewed (although not to the point of significance). For all *wh*-questions except *when* and *who*, the presence of a modal gave rise to a significant MS bias.

Fig. 19 presents several matrix verbs that are of theoretical interest. As in Experiment 1b, we encounter sparse data problems and so focus on *know-wh*. Before turning to those, we observe that with *remember-wh*,

Table 8: Coefficient table (predicted β coefficient, standard error SE, t value, and p value) for *wh*-word -specific models in Experiment 2b. Both predictors are dummy-coded and centered (0 is no Modal Present and *every*-paraphrase, 1 for Modal is present and *a*-paraphrase.

Wh-Word		β	SE	t	p
WHAT	Intercept	.24	.01	21.89	$<$.0001
	ModalPresent	.07	.01	5.76	$<$.0001
	Paraphrase	-.004	.02	-.2	.85
	ModalPresent:ModalPresent	.06	.03	2.5	$<$.02
HOW	Intercept	.02	.01	23.86	$<$.0001
	ModalPresent	.04	.01	3.31	$<$.002
	Paraphrase	.13	.01	11.12	$<$.0001
	ModalPresent:Paraphrase	.1	.02	5.49	$<$.0001
WHERE	Intercept	.2	.02	13.84	$<$.0001
	ModalPresent	.04	.03	1.36	.18
	Paraphrase	.08	.02	3.18	$<$.003
	ModalPresent:Paraphrase	.19	.06	3.23	$<$.002
WHY	Intercept	.18	.01	22.09	$<$.0001
	ModalPresent	.04	.02	2.36	$<$.03
	Paraphrase	.1	.01	7.3	$<$.0001
	ModalPresent:Paraphrase	.12	.03	4.4	$<$.0001
WHO	Intercept	.21	.02	14.13	$<$.0001
	ModalPresent	.05	.06	.97	.34
	Paraphrase	.11	.05	2.18	$<$.05
	ModalPresent:Paraphrase	.05	.1	.51	.61
WHEN	Intercept	.25	.02	10.11	$<$.0001
	ModalPresent	.1	.06	1.73	.1
	Paraphrase	.18	.05	3.39	$<$.004
	ModalPresent:Paraphrase	.17	.12	1.44	.17

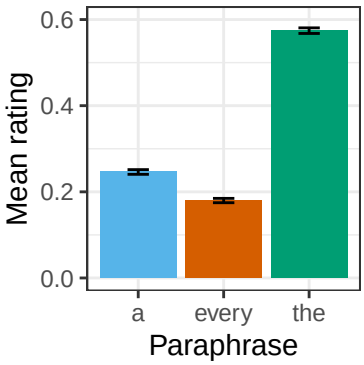

Figure 17: Mean ratings by paraphrase in Experiment 2b. Here and below, error bars indicate 95% bootstrapped confidence intervals.

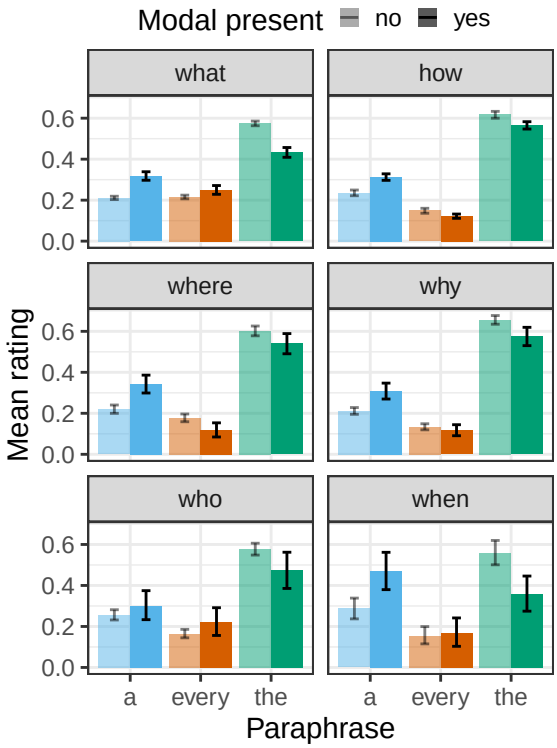

Figure 18: Mean ratings by paraphrase, *wh*-word and modality in Experiment 2b. Error bars indicate 95% bootstrapped confidence intervals.

Fig. 20 presents meaning ratings for *know-wh* questions as a function of paraphrase, Wh-Word and Modality. No *wh*-question showed a bias for MA. Rather we found significant MS bias for *know-how*: β=.16, SE=.02, t=8.97, p <0.0001 and *know-where*: β=.12, SE=.03, t=3.65, p <0.0005; and an MS preference for *know-what*, *know-why*, *know-who*, *know-when*. The MS bias significantly increase with modality for *know-what*: β=.01, SE=.04, t=3.19, p <0.002; *know-how* β=.1, SE=.03, t=3.07, p <0.003, and *know-where* β=.14, SE=.06, t=2.18, p <0.04, but not *know-why*, *know-who*, *know-when*.

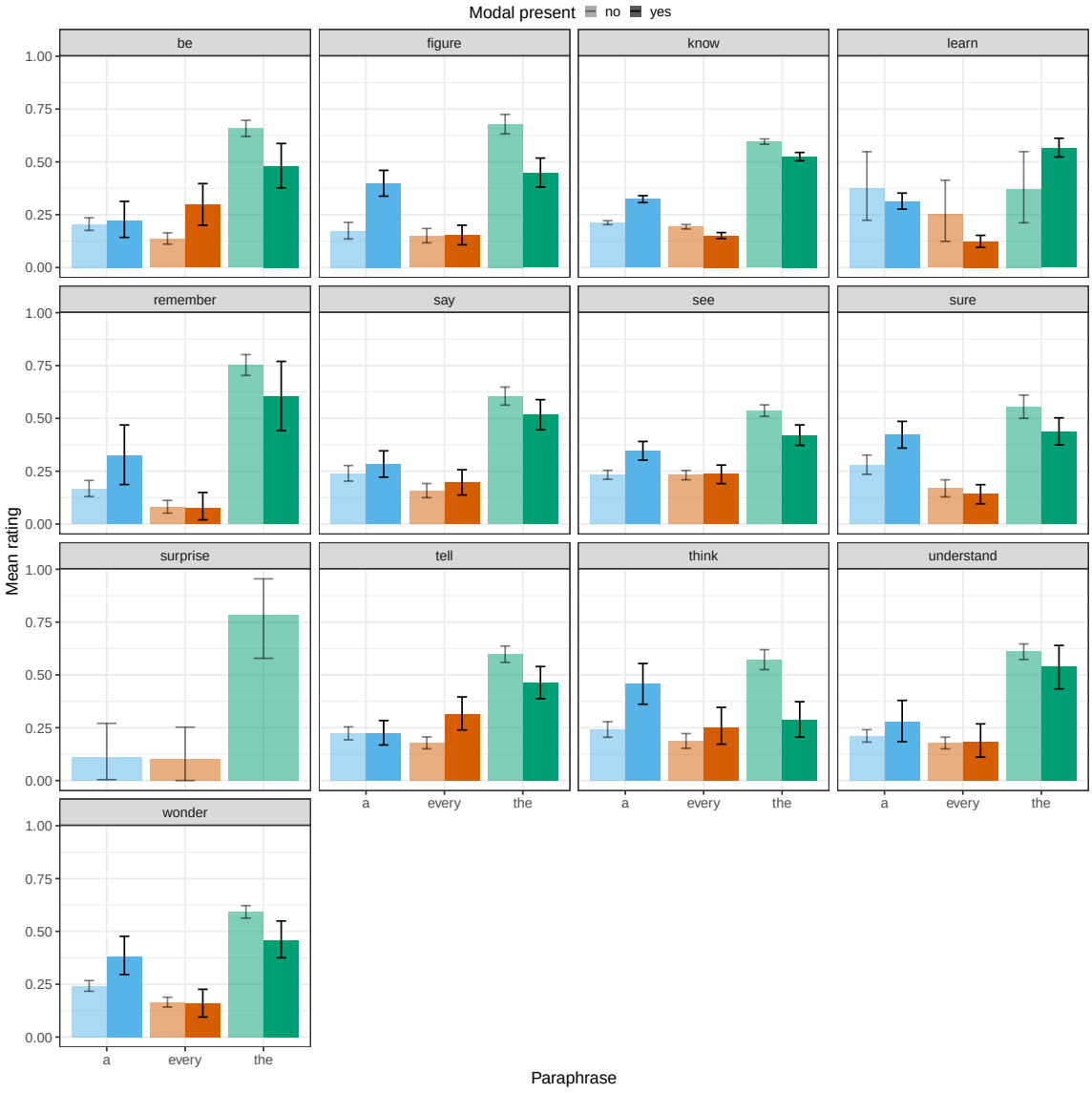

Figure 19: Mean ratings by paraphrase, *wh*-word and modality in Experiment 2b. Error bars indicate 95% bootstrapped confidence intervals.

5.4 Discussion

We found that with the preceeding context removed, both root and embedded *wh*-questions actually showed more bias for MS paraphrases than we expected. Theories which argue that only MS is context dependent would predict that MS readings should decrease without the proper contextual support.

We still found interactions due to the lingusitic form of the question, although there were less robust effects of Modal present, as well as differences due to whether it was the root or the embedded task. For embedded *what*-questions alone the presence of a modal reversed a non-significant MA bias, but there were no higher effects with root questions. For *how*-questions, both root and embedded questions pattern together by ex-

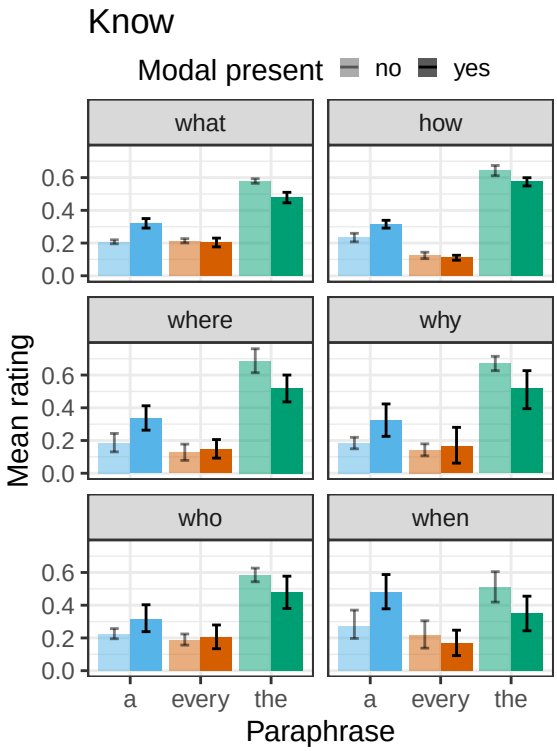

Figure 20: Mean ratings of questions embedded under *know* as a function of paraphrase, *wh*-word and modality in Experiment 2b. Error bars indicate 95% bootstrapped confidence intervals.

hibiting an MS bias that is strengthened by a modal. *Where* and *why*-questions patterned together: for root questions, there was significant MS bias but no influence of modality; for embedded questions the modal did signfifcantly increase the already significant MS bias. For root *who*-questions there was only a significant interaction which raised MS ratings in the presence of modality; for embedded questions, there was significant bias for MS ratings but no effect of or interaction with modality. Finally, *when*-questions showed no interactions between Paraphrase and ModalPresent, but significant positive biases for *a*-paraphrases. Only with Root questions did modality boost MS.

6 Prediction 4: Exploratory meta analysis assessing the effect of context

How crucial is the role of context? In what ways does the presence/absence of explicit linguistic context change the distribution of paraphrases?

If the discourse context provides information about the contextual speaker goals which are crucial to resolving question interpretation, then we should find that interpretation will be harder to resolve in the NoContext condition, where that crucial information is missing. This could manifest in different ways. In the face of this uncertainty about interpretation,

participants might rely more on the linguistic form of the question which provides a cue to the speaker's goal. Thus we might expect to see stronger effects of linguistic factors in the NoContext task than in the Context task. To test this, we introduce Task as a predictor in a regression model over the entire data set, allowing us to quantify the effect of Task and interaction with Wh and ModalPresent. We specifically predict that the effects of Wh and Modality should be stronger in the NoContext Task than the Context Task.

It is also possible participants will be overall more uncertaint about the intended meaning exactly because information about the speaker goal is lacking. This would be reflected in the distributions of paraphrase ratings, where in the NoContext Task the by-item distributions would be more uniform (reflecting that uncertainty) than in Context Task. To test this hypothesis, we compared the by-item distributions for each task to the uniform distribution using the Kullback-Leibler Divergence.

6.1 Qualitative discussion of examples

We subtracted, for each item and paraphrase, the rating in the Context Task from the rating in the NoContext Task. In this section, we discuss some specific examples to help us understand what the effect of context is.

6.1.1 Cases with a big difference between the two studies

For A-paraphrases Let us first discuss examples where mean rating for *a*-paraphrase was higher in the Context Task than in the NoContext Task. In some cases, ratings shifted clearly from *a* in the context experiment to *the* in the NoContext experiment.

(11) a. Speaker A: That's awful.
Speaker B: And he goes out and commits it again. Fact, he's back in jail now. So, what, what, uh,
Speaker A: Gives you sympathy for the vigilantes.
Speaker B: *What deterrent does he really have?*

b. **Rating distributions**:
Context Task [a: .63, every: .16, the: .21]
NoContext Task [a: .18, every: .17, the: .64]

What about the context makes a MS paraphrase better here? For avoiding jail, a single deterrent suffices, but it's not clear from the context that there exists one. Indeed, since the discussion is about recitivism in the relevant person, perphans the speaker is indicating that there exists no deterrent for this person. In contrast, without that information, we don't know what needs to be deterred, nor whether there is a salient deterrent that the speaker has in mind.

In other cases, probability mass shifted from *a* to more evenly between both *every* and *the*:

(12) a. Speaker A: I mean, I'm just thinking of my circle of people that I know, I know quite a few people who have decided to not have both, both, both, uh, couples, you know, both, uh, of the parents work.

Speaker B: yeah.
Speaker A: and yet, uh, I, I we-, I hope to see employer based, you know, helping out. you know, child, uh, care centers at the place of employment and, and things like that, that will help out.
Spealer B: uh-huh.
Speaker A: **What do you think?**

b. **Rating distributions**:
Rating Context Task [a: .43, every: .29, the: .27]
Rating NoContext Task [a: .09, every: .4, the: .5]

The question *What do you think?* is vague. The dialogue in which this question is occurring

In yet other cases, there was a more even spread over the three paraphrases without context. Note that we predicted that removing context would introduce more uncertainty about the intended meanings, refelected in a flatter probability distribution in the NoContext Study.

(13) a. Speaker A: Yeah what did you think about *Dances with Wolves* when you saw it?
Speaker B: Well, okay, see, we're getting back to last year. That's probably the last movie I saw. Um, *Dances with Wolves*, I just adored it.
Speaker A: Really?
Speaker B: **How can I tell you?**

b. **Rating distribution**:
Rating Context [a: .68, every: .03, the: .29]
Rating Nocontext [a: .33, every: .23, the: .45]

Here, context is providing information about what is being told, namely the extent to which Speaker B adored *Dances with Wolves*. Without this information, the question is incredibly vague and it's not clear what the speaker is asking.

Examples where mean rating for *a*-paraphrase was higher without Context In other words, examples where other paraphrases received higher ratings in Contex than out.

Some theories argue that MS requires special contextual licensing. As minimal models of interpretation *all else equal*, these theories predict that MS ratings should be reduced when that special contextual licensing is removed. The items in this section exemplify our finding that in fact, MS ratings increased in the NoContext Task contra theories mentioned above.ß

(14) a. Speaker A: We didn't, we didn't even think about it, you know.
Speaker B: No. And now, you know, what do we have now. You know, got kids that, mumblex either got a, you know, a magnum gun school, like good grief.
Speaker A: R-, right.
Speaker B: I mean, I'd, I'd be afraid to be in school, I mean b-, teaching, or even being a student. **and think what, what's it going to be like for my, my youngest, and my oldest son, when he goes to school?**

b. **Rating Distribution**:
Context Task [a: .14, every: .59, the: .27]
NoContext Task [a: .77, every: .23, the: 0]

This context is about school shootings, a high-stakes context following the criterion of Moyer and Syrett (2019). Presumably, in these kinds of situations, a school will prioritize mapping out as many scenarios as possible so they can be prepared for any school shooter situation. Having exhaustive knowledge is the goal.

(15) a.

b. **Rating Distribution**:
Context Task [a: ., every: ., the: .]
NoContext Task [a: ., every: ., the: .]

(16) a.

b. **Rating Distribution**:
Context Task [a: ., every: ., the: .]
NoContext Task [a: ., every: ., the: .]

For Every-paraphrases In some cases the probability mass shifted between *every* and *the*, and it was the Context task where *every* was rated higher.

(17) a. Speaker A: huh?
Speaker B:
Speaker A: uh-huh.
Speaker B:
Speaker A: I'm sure.
Speaker B: Um, you look at your paycheck **and you go, oh, my gosh where did it all go?**

b. **Rating Distribution**:
Rating Context [a: .07, every: .75, the: .19]
Rating Nocontext [a: .21, every: .21, the: .59]

In this example, the context introduces *paycheck* as the reference for *it*. It is plausible that world knowledge about paychecks is to blame for the raise in the *every* rating here. One doesn't spend one's paycheck (typically) at one single place, i.e., it typicallly is distributed over several places. In this sense, *paycheck* isn't referring to a physical thing, but a collection of money. Note that, without the context, it makes sense that *where did it all go?* would have a high *the* rating, because the speaker appears to be referring to one thing going to one place. [mm: perphase, in part due to definiteness on the pronoun? find some citations like Búring?] The context introduces a more complex antecedent for the pronoun. [mm: It's also interesting to think about the function of *all* in this example.]

(18) a. Speaker A: Well, I, I, I think we did, I think we did learn some lessons that we weren't, uh, we weren't prepared for. I guess the best word would be the atrocities of war.

Speaker B: Yeah.
Speaker A: Uh, I mean the other wars seemed like a valiant war. I mean they seemed like a valiant thing.
Speaker B: Yeah.
Speaker A: **You know, you knew who was good**.

b. **Rating distribution**:
Rating Context [a: .19, every: .6, the: .22]
Rating Nocontext [a: .27, every: .1, the: .63]

There actually weren't many cases where *every* was higher without context. The following is an example of one, but note that the in neither task is *every* rated highest. Rather, *a* is rated highest in the Context Task, and *the* in the No-Context task.

(19) a. Speaker A: and so there's only certain times you can talk to him. and
Speaker B: and you could get there and his office hours could, i mean he could have like a nine to eleven in the morning office hours and have fourty-two people waiting to talk to him, and you still didn't get to talk to him anyway.
Speaker A: right, yeah.
Speaker B: **well, what would be your advice to a parent of a child thinking of attening college?**

b. **Rating distribution**:
Rating Context [a: .44, every: .23, the: .33]
Rating Nocontext [a: .28, every: .48, the: .24]

For The-paraphrases Cases where rating was higher in the Context task:

In most of these cases, participants were highly certain that the *the* paraphrase was the one intended by the speaker in the Context experiment. In the NoContext experiment, they were less certain and we saw probability mass more evenly distributed between the three (although, often *the* was still rated highest).

(20) a. Speaker A: of course I would want one if somebody was given to me.
Speaker B: yeah.
Speaker A: but I maybe would buy a BMW. uh, or, even a Volvo.
Speaker B: uh-huh.
Speaker A:
Speaker B: **What do you have now?**

b. **Rating distribution**:
Rating Context [a: .02, every: .01, the: .96]
Rating Nocontext [a: .25, every: .26, the: .49]

In this example, Context supplies the relevant domain of cars, in which typically people only have a single one. Even though *the* is still highly rated in the NoContext task, note that probability is distributed almost evenly between *a* and *every*.

(21) a. Speaker A: and I did not like giving it out. I mumblex gave out my work number.
Speaker B: Right.

Speaker A: But I think I'm not sure if it's by law just, otherwise I think the practice has basically been eliminated asking for a phone number.
Speaker B: Well, that's the thing I hate too about, uh, radio shack.
Speaker A: **How did radio shack work?**

b. **Rating distribution**:
Rating Context [a: .03, every: .08, the: .89]
Rating Nocontext [a: .33, every: .21, the: .47]

In context, the question is clearly about policies on asking for phone numbers. Furthermore, the existential presupposition that Radio Shack has a policy on the topic is satisfied by Speaker B who explicitly has a (negatively valanced) attitude towards it. Without context, it's less clear which policy is the relevant one (i.e., *work* in what respect?). It's less clear that the presuppositions of *the* are satisfied, reflected in the fact that ratings for *a* and *every* are both higher than in the Context Task (although *the* is still the highest). Participants are less confident that the *the* paraphrase was the intended one.

Cases where rating for *the* was highest in the NoContext task often involved some variation on the question *What do you think?*, which was often rated with an expremely high probability for *the* in the NoContext task (at 1). In the Context task, probability mass was often evenly distributed between all three paraphrases.

(22) a. Speaker A: yeah, that's okay.
Speaker B: Well, that's what I mean like I didn't know what the diffference between Dukakis and Bush was.
Speaker A: Uh-huh.
Speaker B: you know, I didn't know anything about Bush or Dukakis.
Speaker A: **so what do you think about, uh, what do you think about what you see on tv about them, like in the news or on the ads?**

b. **Rating distribution**:
Rating Context [a: .32, every: .36, the: .32]
Rating Nocontext [a: 0, every: 0, the: 1]

(23) a. Speaker A: So long.
Speaker B: Thanks a lot.
Speaker A: Well, uh, ho-, how do you view this whole subject?
Speaker B:
Speaker A: are you, uh, one who feels like you have, have benefited from the change in, in roles in women? **or, or what do you think?**

b. **Rating distribution**:
Rating Context [a: .26, every: .49, the: .24]
Rating Nocontext [a: 0, every: 0, the: 1]

Here's a case where it seems the 10 preceeding lines included the end of one discussion as well as the actual discourse preceeding the target question.

6.1.2 Cases where there was minimal change in rating between the two Tasks

Some cases there was no uncertainty, and no effect of context.

(24) a. Speaker A: I need to make myself do that.
Speaker B: Yeah. I slacked off a little because of, um, I'm about to graduate from college and so this past couple months have been really hectic so I haven't really gone and I've really been faithful these past two months of going to the health club and working out but...
Speaker A: **What school you going to?**

b. **Rating distribution**:
Rating Context [a: 0, every: 0, the: 1]
Rating Nocontext [a: 0, every: 0, the: 1]

(25) a. Speaker A: Generally cheap things.
Speaker B: uh-huh.
Speaker A:
Speaker B: uh-huh
Speaker A:
Speaker B: uh-huh. yeah **we know how that goes**.

b. **Rating distribution**:
Rating Context [a: .18, every: .1, the: .71]
Rating Nocontext [a: .18, every: .11, the: .7]

(26) a. Speaker A: oh, yeah.
Speaker B: and here's tis bum that didn't have a job.
Speaker A: yeah.
Speaker B: and he's got a attorney that you and i could never afford.
Speaker A: that's true.
Speaker B: **Who's paying for that?**

b. **Rating distribution**:
Rating Context [a: .09, every: .07, the: .84]
Rating Nocontext [a: .09, every: .07, the: .84]

6.2 Results

Fig. 21 presents a comparison of the overal results from the Context task compared with the NoContext task. Overall, it seems that the distribution of ratings are largely similar between the two tasks; however, means are numerically higher in the NoContext Task for both *a* (NoContext: mean=0.239, sd=.309, vs. Context: 0.216, 0.331) and *the*-paraphrases (NoContext: mean=0.576, sd=.387, vs. Context: 0.547, 0.432) but the opposite for *every*-paraphrases (Context: 0.237, 0.366, vs. NoContext: mean=0.185, sd=.303).

We can see these differences from a slightly different angle in Fig. 22, which presents the mean by-item ratings for each paraphrase plotted as a function of Task.

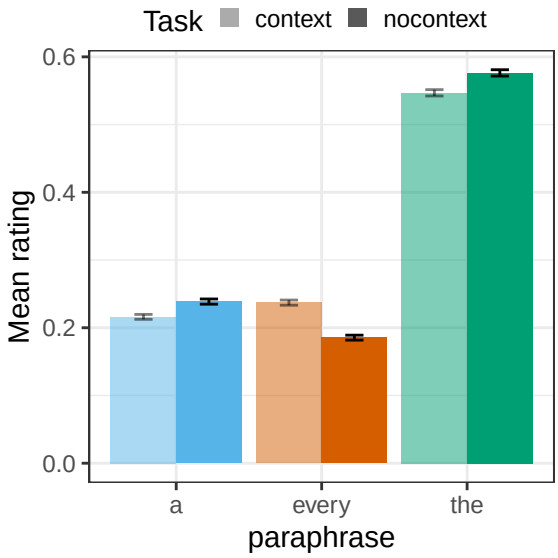

Figure 21: Mean ratings by paraphrase comparing Context and NoContext Experiments.

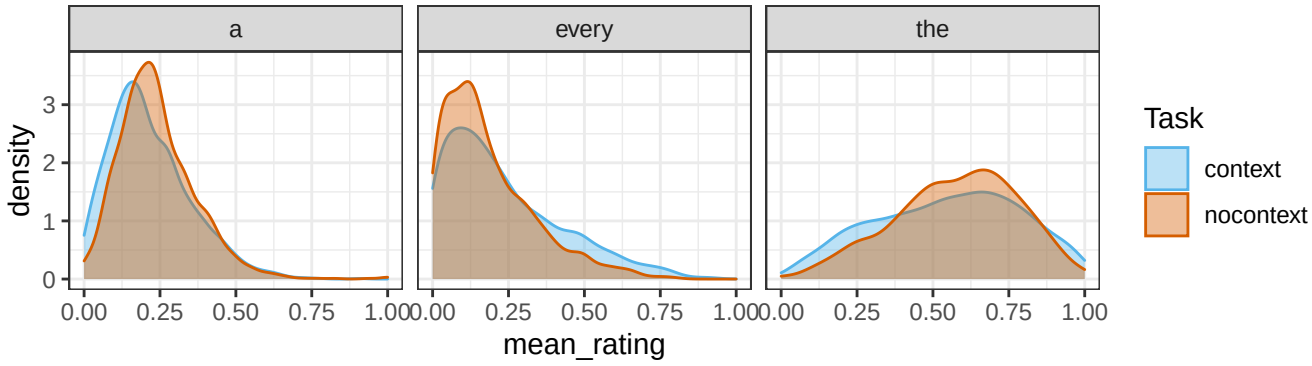

Figure 22: Mean ratings by paraphrase comparing Context and NoContext Experiments.

6.2.1 Regression analysis

Some semantic theories predict that MS only requries special licensing by the context. Thus, if information that plays a licensing role is removed, it is plausible that those theories would predict that MS readings would decrease in the second set of experiments. We already saw that this prediction was not borne out.

Analysis were conducted to assess the effect of Task (Context, NoContext) in addition to modality and *wh*-word on question interpretation. Given the potential 4-way interaction by including Task as a predictor, we first broke the data set up by Paraphrase and then conducted sub-model analysis by *wh*-phrase as previously done for the earlier experiments.

The overall effect of Task on Paraphrase Table 9 presents regression results assessing the effect of Context (Task) on *a*- and *every*- paraphrase ratings. We find first that

Table 9: Coefficient table (predicted β coefficient, standard error SE, t value, and p value) for overall Task x Paraphrase model in metanalysis comparing Context and NoContext tasks. Both predictors are dummy-coded and centered (0 is the Context Task and *every*-paraphrase, 1 for Modal is present and *a*-paraphrase.

	β	SE	t	p
Intercept	.22	.003	80.66	$<$.0001
Paraphrase	.01	.006	2.26	$<$.03
Task	-.01	.003	-4.75	$<$.0001
Paraphrase:Task	.07	.008	8.7	$<$.0001

there is a significant initial bias for *a*- over *every*-paraphrases, indicated by the positive oefficient for Paraphrase, and

First that there is a significant initial bias for *a*-paraphrases

***A*-paraphrase Submodel** Table 10 presents the regression results for the effect of Task and ModalPresent on *a*-paraphrases, split up into submodels for each *wh*-phrase to facilitate model interpretation.

***Every*-paraphrase Submodel** Table 11 presents the regression results for the effect of Task and ModalPresent on *every*-paraphrases, split up into submodels for each *wh*-phrase to facilitate model interpretation.

6.2.2 Quantifying information loss with KL divergence

In addition to comparing We calculated the Kullback-Leibler Divergence from each Task and the uniform distribution over paraphrases to determine whether information about interpretation was lost in the NoContext Task, and conducted a regression analysis to determine whether Task predicted the KL divergence score.

We found that removing the context significantly decreased the KL-divergence score, meaning that ratings in the NoContext Task were closer to the uniform distribution, that information was lost when the context was removed.

6.3 Discussion

We found that effects of Task were significant for Wh-Word, but not for Modality, suggesting that context is informative wrt the wh-domain which importantly helps determine the meaning that the speaker intended. [jd: nice if true]

Table 10: Coefficient table (predicted β coefficient, standard error SE, t value, and p value) for Task x ModalPresent *wh*-word submodels for the data subsetted to *a*-paraphrases. Both predictors are dummy-coded and centered (0 is the Context Task, 1 for Modal is present.

Wh-Word		β	SE	t	p
WHAT	Intercept	.21	.007	31.53	$<$.0001
	ModalPresent	.1	.01	11.15	$<$.0001
	Task	.3	.007	4.37	$<$.0002
	ModalPresent:Task	-.03	.01	-2.86	$<$.005
HOW	Intercept	.25	.006	42.91	$<$.0001
	ModalPresent	.08	.01	6.87	$<$ .0001
	Task	.02	.01	2.31	$<$.05
	ModalPresent:Task	.001	.01	.07	.94
WHERE	Intercept	.22	.02	14.7	$<$.0001
	ModalPresent	.1	.03	3.51	$<$.0006
	Task	.01	.02	.8	.44
	ModalPresent:Task	-.02	.03	-.75	.46
WHY	Intercept	.22	.007	32.84	$<$.0001
	ModalPresent	.09	.02	5.82	$<$.0001
	Task	.02	.01	1.27	.24
	ModalPresent:Task	-.02	.02	-.67	.5
WHO	Intercept	.26	.03	9.12	$<$.0001
	ModalPresent	.12	.04	2.92	$<$.005
	Task	.04	.01	2.87	$<$.04
	ModalPresent:Task	-.03	.04	-.81	.42
WHEN	Intercept	.27	.03	9.25	$<$.0001
	ModalPresent	.22	.07	3.44	$<$.002
	Paraphrase	-.01	.03	-.31	.77
	ModalPresent:Paraphrase	-.01	.06	-.24	.81

Table 11: Coefficient table (predicted β coefficient, standard error SE, t value, and p value) for Task x ModalPresent *wh*-word submodels for the data subsetted to *every*-paraphrases. Both predictors are dummy-coded and centered (0 is the Context Task, 1 for Modal is present.

Wh-Word		β	SE	t	p
WHAT	Intercept	.28	.02	16.09	$<$.0001
	ModalPresent	.05	.02	3.05	$<$.003
	Task	-.08	.01	-6.26	$<$.0001
	ModalPresent:Task	.003	.01	.22	.82
HOW	Intercept	.17	.01	14.83	$<$.0001
	ModalPresent	-.01	.01	-.84	.4
	Task	-.04	.01	-4.02	$<$.0007
	ModalPresent:Task	-.002	.01	-.19	.85
WHERE	Intercept	.19	.02	8.01	$<$.0001
	ModalPresent	-.04	.04	-.97	.34
	Task	-.04	.01	-.422	$<$.0001
	ModalPresent:Task	.03	.03	1.03	.3
WHY	Intercept	.15	.01	17.2	$<$.0001
	ModalPresent	-.001	.02	-.06	.95
	Task	-.03	.01	-2.95	$<$.005
	ModalPresent:Task	-.01	.02	-.62	.54
WHO	Intercept	.23	.02	10.35	$<$.0001
	ModalPresent	.004	.07	.07	.95
	Task	-.09	.02	-4.2	$<$.002
	ModalPresent:Task	-.03	.05	-.49	.62
WHEN	Intercept	.16	.04	4.31	$<$.004
	ModalPresent	.02	.06	.31	.76
	Paraphrase	-.02	.02	-1.01	.28
	ModalPresent:Paraphrase	.03	.04	.71	.48

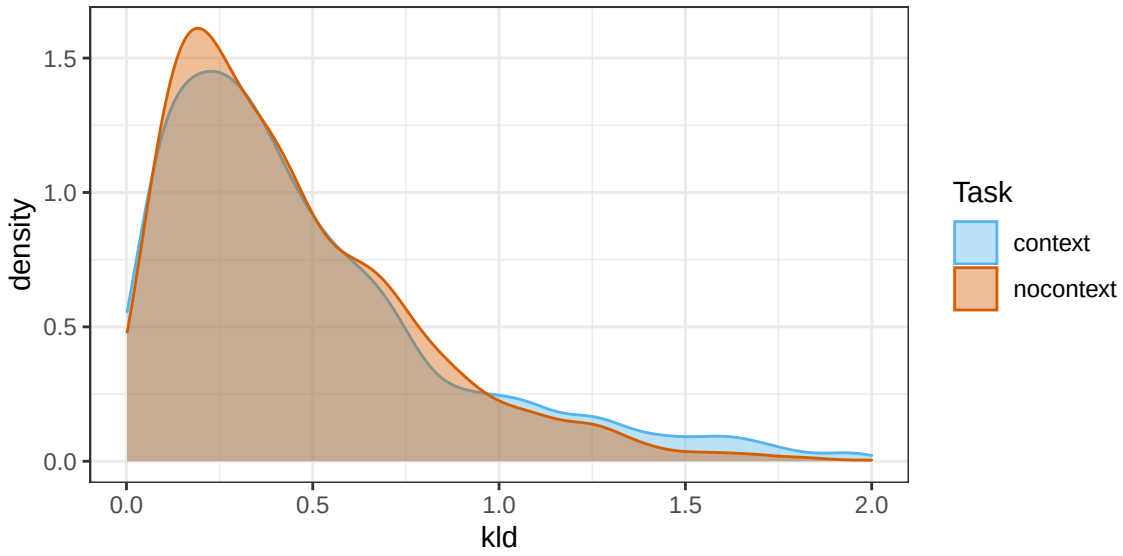

Figure 23: Mean ratings by paraphrase comparing Context and NoContext Experiments.

6.4 Annotation of context for domain-relevant information

We annotated a random sample of contexts of questions that rated high for MS for features that were relevant to determining the size and type of the *wh*-domain. [jd: what is the point of this?]

7 General Discussion

We presented four experiments designed to test the distribution of (non-)exhaustivity in *wh*-questions in naturally occurring speech. Experiments 1a and 2a investigated root questions, while Experiments 1b and 2b investigated questions embedded under propositional attitude verbs [jd: were they all propositional attitude verbs, not embedding predicates in general?]. To test the effect of contextual information on interpretation, Experiments 1a/1b presented targets with the immediately preceding dialogue, while Experiments 2a/2b presented targets without.

[jd: briefly summarize the results here in 3 sentences or so.]

[jd: this next paragraph deserves its own sub-section, should have a heading like "Implications for theories of *wh*-questions ", and contain explicit arguments. you can't just claim that the semantic rep is only existential – you have to say why. you can't just claim that the interpretive process is bayesian – you have to say why (and in fact, we don't have any evidence to that effect from the data presented here, so it's best to just leave this out, or instead frame as a promising direction for how to excplicitly model in-context reasoning about exhaustivity)]

7.1 Implications for semantic theories of

Givn our results, we argue that the semantic representation of a question is maximally underspecified or really weak, existential in meaning. The interpretive process of resolving that underspecification is Bayesian, and involves joint reasoning about the linguistic signal and its possible alternatives, in addition to the contextual speaker's goal.

7.2 Limitations

In our investigation of the linguistic surface features that predict question interpretation, we focused on three that have been widely discussed in the literature: modality, *wh*-word , and matrix verb. We saw that while these are predictive of exhaustivity in ways predicted by previous work [jd: cite cite], in support of claims that variability in exhaustivity is at least to some extent semantically conditioned, we also observed both significant deviations from predictions as well as interactions between factors suggesting that at least some "semantic" effects are in fact defeasible pragmatic ones. [jd: my first stab at this, can be edited of course.]

However, there are other linguistic factors which we did not include in our analyses, and which, if included, would have fully determined the results. [jd: rephrase, jd]

Modality First, we used a coarse-grained measure of modality that included modal auxiliaries and non-finite clauses in embedded questions, but did not account for modal force and flavor ((Kratzer, 1981, 1991; Portner, 2009)). The semantics literature on the effect of modality on MS typically has focused on the existential priority modal *can* as it has occured in the classic example, *Where can I find an Italian newspaper?* On the one hand, we might predict that any modal with existential force would allow for MS because it's the existential bit that drives the MS meaning. This would put the modal observation squarely in line with the observation that existential items in the *wh*-question generally lead to more MS readings than not. It is also concordant with theoretical accounts which argue, either that there is an implicit quantificational variation in the *wh*-question (Hintikka, 1976; Lahiri, 2002), underlying existential quantification that can be strengthened with contextual information (Asher & Lascarides, 1998).

There is a

[jd: this was just getting interesting! :) would love to see where this is going]

In/definiteness Second, definite and indefinite noun phrases may reveal interesting differences with respect to MS and MA. On the one hand, [jd: what has been said about definiteness?]

Other *Wh*-phrases Third, we chose to include only monomorphemic *wh*-phrases, and not include questions with singular or plural marked *wh*-phrases. [jd: remind us why]

Similarly, we did not include clefts or relative clauses. There is a question in the literature on clefts about the extent to which they allow for non-exhaustive readings [mm:

more citations?..check diss, devaug-geiss, zimmermann, destruel]. [jd: i don't think this needs to be listed as a limitation. instead, the connection to variability in cleft/relative clause interpretation can be mentioned in one sentence elsewhere]

Better that pragmatics is not sufficient: [jd: don't understand the point of these examples] (discussion from George 205-206) questions with ostensibly the same semantic representation

(27) a. Who has leprosy?
b. Who are some people with leprosy?

But we can question indeed whether these two questions would be asked by a speaker iwth the same goals in mind?

Question Type [jd: not sure i get what is being said in this section. here i would just focus on the fact that it's hard to actually assess the diffs between root and embedded questions because, with the exception of "know", we have huge data sparisity issues]

Fourth, we did find some differences between root and embedded questions. Embedded questions appeared to slightly prefer MA paraphrases over MS paraphrases, while root questions showed the opposite. This effect could be due to the additional effect of Matrix verb which was not present for Root Questions. Indeed, the majority of embedded questions occured with the verb *know*, which is classically MA. Additionally, there were some cases of embedded questions in which the question contained subject-auxiliary inversion, a connonical root question syntax. Some have argued that MS, being pragmatic, should not be possible in embedded questions ((Karttunen, 1977)[mm: footnote 4], (Groenendijk & Stokhof, 1984)[mm: footnote 14], [mm: xiang?])

Of course, this kind of "pragmatics does not interfere in semantics" view is less in vogue than it used to be, in part because of the growing understanding of pragmatic processing ([mm: reference list from kreiss & degen cog sci paper]). We did an exploratory analysis on the effect of Question Type (true embedded, embedded with subject-aux inversion, true root) and the effects were not overwhelmingly different. [mm: Show graphs? do regression?]

Corpus Genre [jd: why should corpus genre matter? be explicit] The Switchboard corpus is comprised of conversational dyads between randomly assigned employees of Texas Instrunmnets. The conversants were strangers so they did't share a lot of common ground. There are at least two ways in which this affected the corpus. First, it affected the kind of questions asked,

and second, it affected the

Already, the fact that *wh*-question interpretation could be sensitive to all these further factors suggests that

7.3 MISC discussion points

Before continuing, let us pause for a terminological clarification. For root questions, we speak of MS/MA answers, while for embedded questions, we speak of MS/MA readings

or interpretations. The distinction is often theoretically driven. 'Answer' specifically means an element that fills in for the information missing from the *wh*-question . For example, while *I don't know* is a perfectly acceptable response to (1a), it would not count as a answer; in contrast, coffee shop names do fill in the missing *where* information in the right way. So 'answer' is a technical notion.

In the embedded case, we use 'reading' or 'interpretation' because sentences with embedded *wh*-questions do not have interrogative illocutionary force and thus no answers in the speech act sense. Although, semanticists do speak of 'answers' to embedded question, using the technical notion described above. Yet here, 'answers' aren't construed as assertions in the way they are with the root questions. Further, the difference between 'reading' and 'interpretation' is usually theoretically important. 'Reading' implies an underlying semantic representation that determines the truth (answerhood) conditions, while 'interpretation' applies more generally without commitment to an underlying representation. 'Interpretation' often suggests that some other mechanism (usually pragmatic in nature) has intervened to cause the appearance of the data in question. Thus, some might reject the availability of MS *readings* of some embedded questions, but acknowledge the possibility of MS *interpretations* of them, given other (non-semantic) factors. Referring to an MS/MA *reading* of a root question implies that an MS or MA answer suffices as a complete (and resolving) answer to the question. Referring to an MS/MA *interpretation* of a root question implies that a language user has determined that an MS or MA answer is acceptable (though possibly incomplete or unresolving) for a root question, given the exigencies of context. Again, the distinction often boils down to the involvement of non-linguistic processes.

8 Conclusion

[jd: i imagine all the stuff that follows is intended to be used in some form either in the intro or in the gd, but the conclusion should just be short and sweet]

Human communication proceeds remarkably fast and robustly despite the rampant underspecification of speakers' utterances with respect to the meaning they intend to convey. Resolving that underspecification requires that hearers integrate a wide range of possibly uncertain linguistic and extra-linguistic cues.

This view of pragmatics, informed by psycholinguistic research on language processing that espouses a dynamic, nonmodular view of comprehension, has provided a useful novel perspective on many phenomena at the semantics/pragmatics interface. For instance, one of the most-studied cases of underspecification in experimental pragmatics is the scalar inference from *some* to *not all* (e.g., *Scully ate some of the cookies* typically licenses the inference that she did not eat all of them). Recent research using a large dataset of naturally-occurring utterances has revealed a large amount of variability in whether hearers derive scalar inferences (Degen, 2015). Rather than being random, the observed variability was dependent on multiple features of the linguistic and discourse context. Data from controlled experimental tasks with artificially generated stimuli confirm these results: scalar inferences are systematically dependent on (the hearer's estimate of) the speaker's discourse goal (Zondervan, 2010)[mm: kursatdegen2020], the speaker's

epistemic state [mm: goodmanstuhlmueller2013](Breheny, Ferguson, & Katsos, 2013), and which alternatives are contextually available (Huang & Snedeker, 2011; Degen & Tanenhaus, 2016), among other cues. These results were unexpected in light of the theoretical literature, which had predicted a higher prevalence of the inference and no systematic context-dependence (Levinson, 2000).

A good deal of experimental attention has been paid to pragmatic inferences, like scalar inferences, that are the result of reasoning about declarative utterances. In contrast, there is much less discussion about the cues guiding hearers' interpretation of non-declarative utterances like questions. Consider polar (*yes-no*) questions: these can be answered literally with a *yes* or *no*, but often the literal answer is neither the most appropriate, nor what the speaker intended. In Searle (1975)'s classic example, a dinner guest who asks *Can you reach the salt?* likely intends you to pass the salt, not say *yes*. Whether a hearer understands the speaker to want a literal or non-literal answer depends on what they infer about the speaker's goals. Clark (1979) surveyed liquor merchants to determine how they answered a polar question like *Does a fifth of Jim Beam cost more than $5?* when it was introduced by a brief sentence that made the speaker's goal explicit. If the speaker first stated *I want to buy some bourbon*, merchants were more likely to answer with the exact price of the whiskey. If the speaker instead stated *I've got $5 to spend*, merchants were more likely to provide the more literal *yes*/*no* answer. That is, merchants responded by addressing the inferred speaker goal. Research in computational cognitive science has followed suit by modeling question asking and answering as a species of rational, goal-directed behavior (Hawkins, Stuhlmüller, Degen, & Goodman, 2015; Rothe, Lake, & Gureckis, 2018). [mm: hawkinsgoodman2019?]*Wh*-questions are even more complex than polar questions in that even their literal interpretation is underspecified.

References

Asher, N., & Lascarides, A. (1998). Questions in dialogue. *Linguistics and Philosophy*, *21*(3), 237–309.

Beck, S., & Rullmann, H. (1999). A flexible approach to exhaustivity in questions. *Natural Language Semantics*, *7*(3), 249–298.

Berman, S. (1994). *On the semantics of wh-clauses*. New York and London: Garland Publishing, Inc.

Berman, S. R. (1991). *On the semantics and logical form of wh-clauses* (Unpublished doctoral dissertation). University of Massachusetts, Amherst.

Bhatt, R. (1999). *Covert modality in non-finite contexts* (Doctoral dissertation). University of Pennsylvania.

Boër, S. E., & Lycan, W. (1975). Knowing who. *Philosophical Studies*, *28*(5), 299–344.

Breheny, R., Ferguson, H. J., & Katsos, N. (2013). Taking the epistemic step: Toward a model of on-line access to conversational implicatures. *Cognition*, *126*(3), 423–440.

Ciardelli, I., Groenendijk, J., & Roelofsen, F. (2013). Inquisitive semantics: a new notion of meaning. *Language and Linguistics Compass*, *7*, 459–476.

Ciardelli, I., Roelofsen, F., & Theiler, N. (2016). Composing alternatives. *Linguistics and Philosophy*.

Clark, H. H. (1979). Responding to indirect speech acts. *Cognitive psychology*, *11*(4), 430–477.
Comorovski, I. (1996). *The interpretation of interrogative phrases*. Dordrecht, Netherlands: Springer.
Cremers, A., & Chemla, E. (2016). A psycholinguistic study of the exhaustive readings of embedded questions. *Journal of Semantics*, *33*(1), 49–85.
Cremers, A., & Chemla, E. (2017). Experiments on the acceptability and possible readings of questions embedded under emotive-factives. *Natural Language Semantics*, *25*(3), 223–261.
Dayal, V. (1996). *Locality in wh quantification: Questions and relative clauses in hindi*. Dordrecht: Kluwer.
Dayal, V. (2016). *Questions* (C. Barker & C. Kennedy, Eds.). Oxford, United Kingdom: Oxford University Press.
Degen, J. (2015). Investigating the distribution of some (but not all) implicatures using corpora and web-based methods. *Semantics and Pragmatics*, *8*, 11–1.
Degen, J., & Jaeger, T. F. (2011). The TGrep2 database tools. *Unpublished manuscript*.
Degen, J., & Tanenhaus, M. K. (2016). Availability of Alternatives and the Processing of Scalar Implicatures: A Visual World Eye-Tracking Study. *Cognitive Science*, *40*(1), 172–201.
Fox, D. (2014). *Mention-some readings.* MIT seminar notes.
George, B. R. (2011). *Question embedding and the semantics of answers* (Doctoral dissertation). Linguistics Department, University of California, Los Angeles.
Ginzburg, J. (1995). Resolving questions, I & II. *Linguistics and Philosophy*, *18*(5–6), 459–527; 567–609.
Godfrey, J., Hilliman, E., & McDaniel, J. (1992). SWITCHBOARD: A telephone speech corpus for research and development. *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 517–520.
Grimshaw, J. (1979). Complement selection and the lexicon. *Linguistic Inquiry*, *10*(2), 279–326.
Groenendijk, J., & Stokhof, M. (1982). Semantic analysis of wh-complements. *Linguistics and Philosophy*, *5*(2), 175–233.
Groenendijk, J., & Stokhof, M. (1984). *Studies on the semantics of questions and the pragmatics of answers* (Doctoral dissertation). University of Amsterdam.
Hawkins, R., Stuhlmüller, A., Degen, J., & Goodman, N. (2015). Why do you ask? good questions provoke informative answers. In *Proceedings of the 37th annual conference of the cognitive science society.*
Heim, I. (1994). Interrogative semantics and Karttunen's semantics for *know*. In *Proceedings of IATL* (Vol. 1, pp. 128–144).
Hintikka, J. (1974). Questions about questions. *Semantics and philosophy*, 103–158.
Hintikka, J. (1976). *The semantics of questions and the questions of semantics: Case studies in the interrelations of logic, semantics, and syntax*. North-Holland Publishing Company.
Huang, Y. T., & Snedeker, J. (2011). *Logic and conversation* revisited: Evidence for a division between semantic and pragmatic content in real- time language comprehension. *Language and Cognitive Processes*, *26*(8), 1161–1172.

Karttunen, L. (1977). Syntax and semantics of questions. *Linguistics and Philosophy*, *1*(1), 3–44.
Klinedinst, N., & Rothschild, D. (2011). Exhaustivity in questions with non-factives. *Semantics and Pragmatics*, *4*(2), 1–23.
Kratzer, A. (1981). The notional category of modality. In H. J. Eikmeyer & H. Rieser (Eds.), *Words, worlds, and contexts* (pp. 38–74).
Kratzer, A. (1991). Modality. In A. von Stechow & D. Wunderlich (Eds.), *Semantics: An international handbook of contemporary research* (pp. 639–650).
Lahiri, U. (2002). *Questions and answers in embedded contexts*. Oxford University Press on Demand.
Levinson, S. C. (2000). *Presumptive meanings: The theory of generalized conversational implicature*. MIT Press, Cambridge, MA.
Moyer, M. (2020). *The question of questions: resolving (non-) exhaustivity in* wh-*questions* (Doctoral dissertation). Linguistics Department, Rutgers University.
Moyer, M., & Syrett, K. (2019). (Non-)exhaustivity in embedded questions: Contextual, lexical and structural factors. In M. T. Espinal, E. Castroviejo, M. Leonetti, L. McNally, & C. Real-Puigdollers (Eds.), *Proceedings of Sinn und Bedeutung 23* (pp. 207–224).
Nicolae, A. C. (2014). *Any questions? polarity as a window into the structure of questions* (Doctoral dissertation). Linguistics Department, Harvard University.
Pesetsky, D. (1987). Wh-in-situ: Movement and unselective binding. *The representation of (in) definiteness*, *98*, 98–129.
Portner, P. (2009). *Modality*. Oxford: Oxford University Press.
Rohde, D. L. (2005). Tgrep2 User Manual. *Unpublished manuscript*.
Rothe, A., Lake, B. M., & Gureckis, T. M. (2018). Do people ask good questions? *Computational Brain & Behavior*, *1*(1), 69–89.
Searle, J. R. (1975). Indirect speech acts. In *Speech acts* (pp. 59–82). Brill.
Sharvit, Y. (2002). Embedded questions and 'de dicto' readings. *Natural Language Semantics*, *10*(2), 97–123.
Theiler, N. (2014). *A multitude of answers: embedded questions in typed inquisitive semantics* (Unpublished doctoral dissertation). Universiteit van Amsterdam.
Theiler, N., Roelofsen, F., & Aloni, M. (2016). Towards a uniform account of responsive verbs. In *Proceedings of salt 26.*
van Rooij, R. (2003). Questioning to resolve decision problems. *Linguistics and Philosophy*, *26*(6), 727–763.
van Rooij, R. (2004). The utility of mention-some questions. *Research on Language and Computation*, *2*, 401–416.
von Fintel, K. (1994). *Restrictions on quantifier domains* (Unpublished doctoral dissertation). University of Massachusetts, Amherst.
Xiang, Y. (2016). *Interpreting questions with non-exhaustive answers* (Doctoral dissertation). Linguistics Department, Harvard University.
Zondervan, A. (2010). *Scalar implicatures or focus: An experimental approach* (Doctoral dissertation). University of Utrecht.