Assessing the inferential strength of epistemic *must*

Giuseppe Ricciardi, Rachel Ryskin, Edward Gibson

Language, Ahead of Print, (Article)

Published by Linguistic Society of America
DOI: *https://doi.org/10.1353/lan.0.a913402*

ASSESSING THE INFERENTIAL STRENGTH OF EPISTEMIC *MUST*

GIUSEPPE RICCIARDI — *Harvard University*
RACHEL RYSKIN — *UC Merced*
EDWARD GIBSON — *MIT*

This article presents four experiments that investigate the meaning of English and Italian statements containing the epistemic necessity auxiliary verb *must*/*dovere*, a topic of long-standing debate in the philosophical and linguistics literature. Our findings show that the endorsement of such statements in a given scenario depends on the participants' subjective assessment about whether they are convinced that the conclusion suggested by the scenario is true, independently from their objective assessment of the conclusion's likelihood. We interpret these findings as suggesting that English and Italian speakers use epistemic necessity verbs to communicate neither conclusions judged to be necessary (contrary to the prediction of the standard modal logical view) nor conclusions judged to be highly probable (contrary to the prediction of recent analyses using probabilistic models) but conclusions whose truth they believe in (as predicted by the analysis of epistemic *must* as an inferential evidential). We suggest that this evidential meaning of epistemic *must*/*dovere* might have arisen in everyday conversation from a reiterated hyperbolic use of the words with their original meaning as epistemic necessity verbs.*

Keywords: semantics, epistemic modality, evidentiality, experimental methodology, inductive vs. explanatory inference, necessity, probability

1. INTRODUCTION. In our everyday life we often form beliefs about events starting from a state of uncertainty and relying on our reasoning. Beginning with Tversky & Kahneman 1974, studies in the psychology of decision making and reasoning have investigated the cognitive mechanisms underlying how we reason to form such beliefs under conditions of uncertainty (e.g. Tversky & Kahneman 1992, Gigerenzer et al. 2011, Preuschoff et al. 2013, Gonzalez 2016). Overall, this line of work has shown that people tend to violate laws of logic or probability in determining their confidence in the possible occurrence of an event and rely, instead, on heuristics that simplify the likelihood-estimation task. These heuristics are based on cognitive biases, which result in people adopting beliefs about the occurrence of an event based on a subjective probability of that event more or less independently of its objective probability.

A well-known case of these cognitive biases is the so-called EXPLANATION EFFECT, such that 'an explanation's quality is used as a guide to the probability of that explanation' (Lombrozo 2012:269; see also Chapman & Chapman 1967, 1969, Ross et al. 1977, Anderson et al. 1980, Koehler 1991, Brem & Rips 2000, Lombrozo 2007). For example, Lombrozo 2007 shows that an explanation's degree of simplicity is used as a cue for its likelihood: when participants learned about a patient with two symptoms, they overwhelmingly considered it to be more likely that the symptoms were caused by a single disease (simple explanation) than by the conjunction of two diseases (complex explanation) in the absence of base rates for the diseases. Other properties of explanations that have been shown to increase their estimated likelihood are their breadth, coherence, and consistency with prior knowledge (see Lombrozo 2012 for an overview of this literature).

* First, we would like to thank Kristen Syrett and three anonymous referees for *Language* for their detailed comments and suggestions, which have strongly improved the original manuscript. For their valuable feedback at various stages of the project, we also thank Ray Jackendoff, Manfred Krifka, Kathryn Davidson, Daniel Lassiter, Brandon Waldon, Susi Wurmbrand, Deniz Rudin, Gennaro Chierchia, Shannon Bryant, Joshua Martin, the members of TedLab at MIT, the members of the Meaning and Modality Lab at Harvard, and the audiences of LSA 2019 and ELM 2020. This research was supported by a grant from the National Science Foundation (Award 2121074) to E. Gibson and R. Levy.

Human languages offer many expressions that modulate our degree of confidence about the occurrence of an event, such as *I think that …* , *I guess that …* , *probably …* , *it is certain that …*, *necessarily …* , *I know that …* , *perhaps …* , and so forth. However, not much work has been done to connect the investigations conducted by psychologists into how humans reason under uncertainty with those conducted by linguists into the meaning of expressions describing the output of that reasoning process.

In this project, we aim to fill this gap by showing that a proper understanding of how people form beliefs under conditions of uncertainty can help shed light on the meaning of the 'necessity auxiliary verbs' in their epistemic interpretation. Specifically, we focus on the epistemic interpretation of the English verb *must* and the Italian verb *dovere*, as exemplified in 1 and 2, respectively, with corpus-searched instances.

(1) *Must* in its epistemic interpretation
   a. When you say that your students learned less, you **must** have had some mechanism to measure that, right?[1]
   b. Everyone says that. I guess it **must** be true.[2]

(2) *Dovere* in its epistemic interpretation
   a. **Deve** essere stato un brutto incidente perché sento la sirena della polizia.[3]
      'It must have been a bad accident because I hear the police siren.'
   b. Lavorare con tre bambini piccoli non **deve** essere stato facile agli inizi.[4]
      'To work with three kids must not have been easy at the beginning.'

The meaning of epistemic necessity auxiliaries like *must* and *dovere* has been a matter of lively debate in the semantic and philosophical literature since at least the 1970s (see, among others, Jackendoff 1972, Karttunen 1972, Lyons 1977, Kratzer 1981, 1991, Stone 1994, Westmoreland 1998, Drubig 2001, Palmer 2001, Papafragou 2006, Stephenson 2007, von Fintel & Gillies 2010, 2021, Giannakidou & Mari 2016, 2018, Lassiter 2016, 2017, Goodhue 2017, Degen et al. 2019, Del Pinal & Waldon 2019, Mandelkern 2019). Logically oriented semanticists start from the observation that the words *must* and *dovere* (as well as other modal verbs like *may*, *should*, *might*, etc.) are ambiguous between different meanings, whose exact number and definitions are debated (see Nuyts 2016 for a critical overview of the proposed classifications). However, there is agreement that at least two types of meaning should be identified for modal verbs: the deontic meaning (as in *Mary must wash the dishes*) and the epistemic meaning (as exemplified in 1). Next, they assume that all of these senses share a common basic meaning corresponding to the meaning of the necessity operator in modal logic, which can be roughly paraphrased as 'it is necessary that/necessarily' (Hintikka 1962, Carnap 1964, Lewis 1968, Kratzer 1981). Lastly, they hypothesize that the ambiguity between deontic and epistemic sense (and all of the other potential senses) arises from the assumptions on which the claim 'it is necessary that/necessarily' is based. Specifically, a statement containing a necessity auxiliary like *must* is epistemic when it is used to evaluate an event as epistemically necessary in view of some known facts—that is, as maximally likely (100% certain) based on those facts (which can be explicitly mentioned by the speaker or left implicit). For example, according to this view, the

[1] Source: http://blogs.chicagotribune.com/news_columnists_ezorn/2012/09/test-anxiety.html

[2] Source: FIC: Analog Science Fiction $26 Fact; Date: 2019; Publication information: Vol. 138, Iss. 9/10; Title: IMPETUS.

[3] Source: http://notimetolose.myblog.it/tag/morte

[4] Source: http://campodarsego.blogolandia.it/2009/10/28/campodarsego-class-intervista-il-presidente-roncato-donna-del-fare-forse-troppo-visto-che-il-consiglio-gli-ha-sbattuto-la-porta-in-faccia/

statement in 1a can be paraphrased as follows: 'When you say that your students learned less, **it is necessarily the case that** you have had some mechanism to measure that, right?'. We refer to this account as the LOGICAL *MUST* HYPOTHESIS. For simplicity, we use the label '*must p*' as a cover term for the class of statements containing necessity auxiliary verbs in their epistemic interpretation.

(3) LOGICAL *MUST* HYPOTHESIS: *must p* = 'Given some salient facts, the likelihood of *p* is maximal.'

Crucially, this analysis yields the prediction that in saying *must p* English speakers are committing themselves to take *p* as an established fact, thereby committing themselves to also assert *I know that p* and *It is certain that p*.

Such predictions, however, seem to be at odds with the intuition that was first expressed in Karttunen 1972:12: 'In general one would use [*must p*] only in circumstances where it is not yet an established fact that [*p*]'. More recently, Lassiter (2016:121) has supported this intuition with several corpus examples suggesting that English '[s]peakers use *must p* when they are not maximally certain of *p*; when they explicitly consider ¬*p* to be a possibility; and when their stated grounds for concluding that *p* must be the case are less than fully compelling, and even explicitly stated to be so by the speaker'. The debate about the strength of epistemic *must* is defined by attempts to answer two questions: (i) Should the logical *must* hypothesis be rejected because it is too strong to account for the actual use of *must p*? (ii) If it should be rejected, then what is the meaning of epistemic *must*? Within this debate, we have identified five positions, which we present in the order of how conservative they are relative to the logical analysis.

The most conservative hypothesis has been advanced by von Fintel and Gillies (2010, 2021). They argue that the modal logic analysis defines well the literal meaning of *must p*: a speaker uttering *must p* is communicating 'Given some salient facts, *p*'s likelihood is maximal', thereby committing themself to also assert *I know that p* and *I'm*/*It's certain that p*. In order to account for the perceived weakness of these statements, von Fintel and Gillies (2021) suggest that *must p* tends to be used hyperbolically—namely, in situations where *p* is very close to being certain based on salient facts but, strictly speaking, is just highly probable—which may give the impression that *must p* is weaker than it actually is. In other words, the authors propose that cases where epistemic *must* is used to talk about a very probable but not certain event should be analyzed in analogy to cases where we say, for example, 'It's 3 PM' but actually it is 2:59 PM: by saying 'It's 3 PM' we said something that is strictly speaking false but easily tolerable in casual conversations, where the exact time is not crucial. Similarly, when a speaker utters, for example, 'Susan must be in her office. I see the light on', they are exaggerating in presenting the event of Susan being in her office as certain given the fact that the light is on inside the office, disregarding the possibility that she might have left the office and forgotten that the light was on. However, as soon as someone points this possibility out, the speaker has to admit that 'Susan must be in her office because the light is on in her office' is false as much as 'It is certain that Susan is in her office because the light is on in her office'. So, in summary, von Fintel and Gillies speculate that *must p* feels not as strong as predicted by the standard logical analysis because people typically overuse it in cases where more humble expressions like *probably p* would be more appropriate. We refer to this proposal as the HYPERBOLIC LOGICAL *MUST* HYPOTHESIS.

(4) HYPERBOLIC LOGICAL *MUST* HYPOTHESIS: *must p* = 'Given some salient facts, the likelihood of *p* is maximal', *and*
Speakers tend to use this claim in an exaggerated fashion.

Another hypothesis was introduced in Kratzer 1981 (see also Kratzer 1991, 2012) and since then has been endorsed and refined by several other scholars (Veltman 1996, Giannakidou & Mari 2016, 2018, Goodhue 2017, Del Pinal & Waldon 2019, Del Pinal 2021). This proposal maintains from the standard logical analysis the assumption that epistemic *must* is associated with the concept of the maximal likelihood of an event but denies that this likelihood estimation is relativized to facts only: '[i]n uttering [*must p*] … I signalize that I don't reason from established facts alone. I use other sources of information which may be more or less reliable' (Kratzer 1981:57). What are these other sources of information? Kratzer refers to them as 'normality assumptions' or 'stereotypical beliefs', that is, beliefs describing reasonable expectations about what is a normal course of events. According to this hypothesis, in asserting *must p* speakers communicate that the likelihood of *p* is maximal given some contextually salient facts and some contextually salient normality assumptions they make. For example, in asserting *Susan must be in her office. I see the light on*, a speaker is communicating that they judge the event of Susan being in her office to be certain if one assumes that the light in her office is on and that if the light is on in one's office, one is inside. So, in summary, Kratzer speculates that the standard logical analysis is not wrong in postulating that epistemic *must* expresses the maximal (strongest) degree on an event's likelihood scale, but is wrong in postulating that this maximal likelihood estimation is relativized to facts only: people include their beliefs among the assumptions relative to which they judge an event as certain. We refer to this as the WEAK LOGICAL *MUST* HYPOTHESIS.

(5) WEAK LOGICAL *MUST* HYPOTHESIS: *must p* = 'Given some salient facts and normality beliefs, the likelihood of *p* is maximal.'

A third hypothesis was advanced in Swanson 2011 and re-elaborated in Lassiter 2016, 2017. This proposal departs more radically than the first two from the standard logical analysis because it abandons the critical assumption held by logicians that epistemic *must* is linked to the concept of maximal likelihood. In fact, according to this hypothesis, in asserting *must p* speakers communicate that they judge *p* to be a highly probable event given some salient facts. For example, in asserting *Susan must be in her office. I see the light on*, a speaker is communicating that they judge it to be highly probable that Susan is in her office now given the fact that the light in her office is on now. We refer to this proposal as the PROBABILISTIC *MUST* HYPOTHESIS.

(6) PROBABILISTIC *MUST* HYPOTHESIS: *must p* = 'Given some salient facts, the probability of *p* is very high.'

A fourth hypothesis was advanced in Stone 1994, Westmoreland 1998, and Drubig 2001, and more recently defended in Mandelkern 2019. According to this proposal, epistemic *must* is an inferential marker, that is, an expression indicating that the speaker's source of information for the claim is an inference. But what is an inferential marker? In the formal semantic literature, there are several proposals about the meaning of inferential markers and evidential markers more generally (see, among others, Izvorski 1997, Faller 2002, 2019, Matthewson et al. 2007, McCready 2010, Murray 2017). Leaving aside differences in analytical details across the proposals, they all share the following core meaning: a statement containing an inferential marker is used to communicate a piece of information that the speaker (or some other relevant agent in embedding contexts) has acquired through an act of reasoning. Thus, according to this hypothesis, the communicative import of (matrix) *must p* is roughly equivalent to the communicative import of statements containing an attitude verb of inference in the first person like

I conclude that p or *I deduce that p* or *I infer that p*.[5] For example, in asserting *Susan must be in her office. I see the light on*, a speaker is communicating that they have concluded that Susan is in her office from the fact that the light in her office is on. In summary, this proposal assumes that people say *must p* in situations where some salient evidence has made them draw the inference that the event described by *p* has occurred or is occurring. We refer to this proposal as the EVIDENTIAL *MUST* HYPOTHESIS.

(7) EVIDENTIAL *MUST* HYPOTHESIS: *must p* = 'Given some salient facts, the speaker concludes that *p*.'

A fifth hypothesis was advanced in Lyons 1977. Lyons suggests that epistemic *must* is polysemous between two senses. In one sense, 'the English verb *must* has the same function as the modal operator of logical necessity' (Lyons 1977:789), which he refers to as the 'objective epistemic *must*'. In the other sense *must* has a meaning that can be paraphrased as 'I (confidently) infer that', which he refers to as the 'subjective epistemic *must*'. Straightforwardly, the objective epistemic *must* corresponds to the epistemic *must* under the standard logical hypothesis, whereas the subjective epistemic *must* corresponds to the epistemic *must* under the evidential hypothesis. Though Lyons suggests that, in principle, the two meanings are available, he also remarks that the subjective epistemic (evidential) *must* 'in the everyday use of language is of more frequent occurrence' (1977:798). Based on this, we think we are justified in subsuming Lyons's proposal under the label of the evidential *must* hypothesis, with the awareness that with this hypothesis he does not rule out the standard logical hypothesis as an accurate explanation of some less common uses of epistemic *must*.

In fact, we think it is fair to assume that none of the scholars mentioned would deny that the standard logical hypothesis is basically right in assuming that *must* and *dovere* are originally associated with the concept of 'necessity' (in the deontic realm at least, it is undisputed that the use of *must* signals morally necessary events). However, at the same time, it is also intuitively undeniable that, in their typical epistemic uses, the words seem to convey a less strong meaning than that of an operator expressing maximal likelihood of an event given other facts.[6] So, the critical question is: which of the four hypotheses—hyperbolic logical, weak logical, probabilistic, or evidential—provides a better account of what people are communicating in their typical epistemic uses of *must* and *dovere*? Are people communicating in an exaggerated fashion that an event is certain based on some factual evidence (as assumed by the hyperbolic logical hypothesis)? Or are they communicating that an event is certain based on some factual evidence and some stereotypical beliefs (as assumed by the weak logical hypothesis)? Or that an event is highly probable based on some factual evidence (as assumed by the probabilistic hypothesis)? Or that they have reached the conclusion that an event happened based on some factual evidence (as assumed by the evidential hypothesis)?

DIFFERENTIAL PREDICTIONS AMONG THE FOUR HYPOTHESES. Here, we evaluate the meaning of *must*/*dovere* by assessing English and Italian speakers' behavior in a comprehension task requiring participants to decide whether they endorse a statement based

[5] We leave aside here the important question of whether the semantic contribution of evidentials is at the at-issue or not-at-issue level.

[6] This intuition can be easily corroborated by looking at naturally occurring examples like those reported in Lassiter 2016.

on the given contextual information. Let us consider how these four hypotheses differ in terms of their predictions about the behavior of comprehenders in this task.

We take the hyperbolic logical *must* hypothesis as predicting that speakers would endorse *must p* only in contexts where the given information prompts them to also endorse *it is certain that p* or *I know that p*. But this hypothesis is also consistent with people sometimes exaggerating (generating hyperbole) in some contexts.

We take both the probabilistic *must* hypothesis and the weak logical *must* hypothesis as predicting that speakers would endorse *must p* in every context where the given information prompts them to also endorse *it is highly probable that p*. In other words, according to these two hypotheses, for a speaker to judge *p* as highly probable is a sufficient condition for endorsing *must p*.[7]

To understand the predictions of the evidential *must* hypothesis, it is crucial to understand the meaning of *I conclude*/*deduce*/*infer that p*. And to do so, one needs to examine the concept of inference. Inferences can be divided into two types: logical (the inferred conclusion is necessarily true if all of the premises are true) and nonlogical (the inferred conclusion could be false even if all of the premises are true). Assuming this categorization, there exists one type of inference—logical inference based on true premises—that entitles the agent who draws the inference to claim that the conclusion describes a fact, that is, that the conclusion represents a piece of knowledge. Thus, in principle, it is possible to be in a situation where a piece of information obtained through an inference counts as knowledge. But in daily life there is not much room for logical inferences; people are therefore biased to consider inferred propositions as describing opinions of the agent who draws the inference. As a consequence, people are biased to interpret someone uttering 'I conclude/deduce/infer that *p* given the set of facts A' as communicating to us 'I formed the belief that *p* given the set of facts A'. So, under the evidential hypothesis, *must p* is as strong as *I formed the belief that p based on some relevant evidence*.

Assuming this interpretation of the evidential hypothesis, we proceed now to show how the evidential hypothesis makes different predictions from the hyperbolic and the probabilistic hypotheses in endorsement tasks. On the one hand, one can believe a conclusion without judging that conclusion to be certain; that is, judging a conclusion to be certain is not a necessary condition for believing that such a conclusion is true. So, the evidential hypothesis, unlike the hyperbolic logical hypothesis, predicts that speakers can endorse *must p* in contexts where they do not endorse *it is certain that p* or *I know that p*. On the other hand, judging a conclusion as probable is not a sufficient condition for believing in its truth (if one judges *p* to be probable, one does not automatically believe *p*). For example, consider the issue of whether there is life in the universe other than on Earth. Based on what experts say, it is very likely that there is, and many people

[7] A referee pointed us to an alternative interpretation of the weak logical *must* hypothesis described in Del Pinal & Waldon 2019 and Del Pinal 2021. In this interpretation, the weak logical hypothesis differs from the probabilistic hypothesis: the weak logical view predicts that *must p* commits its speakers to believe *p*, whereas the probabilistic view 'only commits [the speaker] to believing that [*p*] has a high likelihood' (Del Pinal & Waldon 2019:158). Del Pinal and Waldon's version of weak *must* is empirically indistinguishable from the evidential *must* hypothesis, because both hypotheses crucially imply that a speaker uttering *must p* is committed to 'I believe that *p* (based on some reasoning)'. Determining which of these two interpretations of the weak *must* hypothesis is meant by the original authors is not our goal here; we are interested in establishing whether this hypothesis makes different predictions from the other three hypotheses. Regardless of whether we adopt our interpretation or Del Pinal and Waldon's, the weak logical *must* hypothesis makes predictions that are identical to those made by one of the other three hypotheses.

would endorse this objective assessment. However, would the same people be ready to claim *I think that there is life in the universe other than on Earth*? Not necessarily. In fact, the mental act of judging an event to be very likely—although it seems quite strong—is actually weaker than the mental act of believing that the event happened: by believing a conclusion one is undertaking a COMMITMENT TO THE TRUTH OF THAT CONCLUSION, which is not the case when one deems the conclusion's likelihood to be high. The evidential hypothesis, unlike the probabilistic *must* hypothesis, thus predicts that speakers can decide to not endorse *must p* in contexts where they endorse *it is highly probable that p*. Overall, the evidential hypothesis predicts that the endorsement of *must p*—as well as statements containing inferential attitude verbs (e.g. *conclude*) or inferential evidentials—is determined by speakers' subjective assessment of whether they are convinced of the truth of a conclusion suggested by the relevant evidence, independently of their objective assessment of the likelihood of that conclusion given the evidence.

In summary, we have identified three hypotheses about the weak common use of *must p*, which make different predictions in an endorsement task, as summarized in Table 1.

HYPOTHESES	PREDICTIONS
HYPERBOLIC LOGICAL *MUST*	Speakers would endorse *must p* only in contexts where the given information prompts them to judge *p* as certain (with some expected exceptions due to exaggerated uses).
PROBABILISTIC *MUST* (= WEAK LOGICAL *MUST*, for us)	Speakers would endorse *must p* in every context where the given information prompts them to judge *p* as highly probable.
EVIDENTIAL *MUST* (= WEAK LOGICAL *MUST*, for Del Pinal & Waldon)	Speakers would endorse *must p* in every context where the given information prompts them to conclude (= form the belief) that *p*.

TABLE 1. The three hypotheses assessed in this work, with their predictions in comprehension tasks prompting participants to decide whether they endorse a statement based on the given information.

The debate about the strength of epistemic *must* has been based primarily on evidence from authors' intuitions, but recently a few studies have aimed at experimentally assessing the hypotheses under discussion (Lassiter 2016, Degen et al. 2019, Del Pinal & Waldon 2019). In particular, the experiment reported in Lassiter 2016 represents the first attempt to test these hypotheses in a comprehension task across many participants. We review this experiment immediately below and postpone the discussion of other relevant findings to the general discussion.

Participants in Lassiter 2016 were provided with a lottery scenario in which the probability of the event of Bill having won the lottery is known and is very small (one chance out of 1,000), and they judged whether they agreed or disagreed with a single statement from a list of nine (see 8), including *Bill must not have won the raffle* ('must not'), *It is certain that Bill did not win the raffle* ('certain not'), and *We know that Bill did not win the raffle* ('know not').

(8) Materials and summary of results from Lassiter 2016
Lottery scenario: Yesterday, Bill bought a single ticket in a raffle with 1000 total tickets. There were also 999 other people who bought one ticket each. That is, the tickets were distributed like this: People holding one ticket: Bill, Mary, Jane, … [997 more]. The drawing was held last night, and the winner will be announced this evening.

List of sentences		
a. Bill won the raffle.	(*did*)	7%
b. Bill did not win the raffle.	(***did not***)	69%
c. It is possible that Bill won the raffle.	(*possible*)	92%
d. Bill possibly won the raffle.	(*possibly*)	74%
e. **We know that Bill did not win the raffle.**	**(*know not*)**	**21%**
f. **It is certain that Bill did not win the raffle.**	**(*certain not*)**	**25%**
g. Bill certainly did not win the raffle.	(*certainly not*)	54%
h. Bill might have won the raffle.	(*might*)	80%
i. **Bill must not have won the raffle.**	**(*must not*)**	**58%**

Lassiter's (2016) main findings were: (i) a majority of participants (58%) agreed with *must not* in the lottery scenario, and (ii) the proportion of participants who agreed with *must not* (58%) was significantly higher than the proportion who agreed with *certain not* (25%) and *know not* (21%)—suggesting that the *it is certain that p* and *we know that p* statements are evaluated as expressing a stronger confidence in the truth of *p* than the *must p* statement. Lassiter took these findings as supporting the probabilistic *must* hypothesis over the logical *must* hypothesis for English epistemic *must*.

However, we find Lassiter's (2016) interpretation of these findings unsatisfactory, because, although the probabilistic *must* hypothesis can account for the behavior of the 58% of participants who agreed with the statement, it does not have an explanation for the behavior of the 42% who disagreed with it. The other two hypotheses described above, by contrast, can account for the behavior of both groups. According to the hyperbolic logical *must* hypothesis, those who agreed with *must not* in the context exaggerated in judging as certain an event that is only probable; by contrast, those who did not agree with *must not* in the context were interpreting the statement with its literal meaning. According to the evidential *must* hypothesis, those who agreed with *must not* in the context judged the provided information as sufficient to believe that Bill did not win the lottery without judging such an event as necessary; by contrast, those who did not agree with *must not* in the context were more cautious and did not want to jump to such a conclusion. Thus, Lassiter's (2016) findings confirm the intuition that the logical *must* hypothesis is too strong but do not discriminate among the other three hypotheses reviewed above.

In this article, we offer findings from an attempted replication of Lassiter 2016 (experiment 1), two follow-up studies with English speakers (experiments 2, 3), and one follow-up study with Italian speakers (experiment 4) in which we manipulated the original task to discriminate among the three hypotheses reviewed above. Overall, our findings support the evidential hypothesis over the hyperbolic logical hypothesis and the probabilistic hypothesis for both English *must* and Italian *dovere* and provide further confirmation for the findings in the psychology of decision making that people form the belief that an uncertain event happened by relying more on the subjective probability of that event than on its objective probability. Moreover, in the general discussion, we speculate that this evidential sense of *must* and *dovere* is a derivative meaning of the words stemming from their overuse as markers of epistemic necessity: the original meaning indicating logical conclusions (i.e. certainties) is weakened/bleached to indicate just conclusions (i.e. opinions) after reiterated and implausible exaggerated uses of the words in their logical sense. So, we think that the hyperbolic logical hypothesis is right in identifying a process of exaggeration as the key component in the epistemic *must* puzzle, but that it is wrong in assuming that, at the current stage of the language, speakers exaggerate when they use the epistemic *must*. In fact, following reiterated exaggerated uses as necessity operators, *must* and *dovere* in their epistemic uses have turned into inferential

markers: in using them, speakers typically are not communicating exaggerated confidence in the certainty of an event but simply that based on their reasoning they have formed the belief that the event happened.

2. EXPERIMENTS.

2.1. EXPERIMENT 1: REPLICATION OF LASSITER 2016. In experiment 1 we attempted to replicate Lassiter 2016. We focused on three conditions from the original nine—'must not', 'know not', and 'certain not'—because these are the conditions relevant for Lassiter's primary conclusions. The materials, data, and statistical analyses relevant to this experiment can be found in the OSF project https://osf.io/ukp2w/.

METHODS.

Participants. We recruited 180 Amazon Mechanical Turkers (sixty for each sentence).

Materials and design. The three critical sentences are given in 4 (unlike Lassiter 2016, we labeled them without 'not').

(9) Experiment 1 sentences (between-subjects design)
   a. Bill must not have won the raffle. (*must*)
   b. It is certain that Bill did not win the raffle. (*certain*)
   c. We know that Bill did not win the raffle. (*know*)

The experiment used a between-subjects design. Each participant saw one critical sentence/trial. Participants read instructions, followed by the target sentence, a radio-button choice between 'agree' and 'disagree', and a simple yes/no question (intended to weed out participants who might not read the context carefully). An example trial is displayed in 10.

(10) Sample trial: experiment 1
Please read the context and the sentence, state whether you agree or disagree with the sentence in the context and then answer the question immediately following.
Context: [The same lottery scenario as in Lassiter 2016]
Target sentence: Bill must not have won the raffle.
Agree Disagree
Question: Is there anyone other than Bill who bought a ticket?
Yes No

PREDICTIONS. The dependent measure was the proportion of 'agree' choices for each sentence. Because this experiment was designed as a replication of a subset of Lassiter 2016, we compared only the two theories considered there: the logical *must* hypothesis and the probabilistic *must* hypothesis. We consider the other theories in the discussion and in later experiments. The logical *must* hypothesis predicts that participants will not agree with the *must* statement, nor the *certain* or *know* statements: the 'agree' proportions should be close to zero for all three. In contrast, the probabilistic *must* hypothesis predicts that participants will agree with the *must* statement (the 'agree' proportion should be close to 1), more so than for the *certain* and *know* statements.

RESULTS. We excluded data from twenty-five participants because they did not fit all of the following inclusion criteria: (a) indicating English as their native language and (b) the USA as country of origin, (c) giving a correct answer to the sanity-check question *Is there anyone other than Bill who bought a ticket?*, and (d) participating in only one condition. This left 155 participants. The number of data points, mean agreement rate, and standard deviation for each of the three sentences are reported in Table 2. The mean agreement ratings with error bars are plotted in Figure 1 (middle panel).

SENTENCE TYPE	COUNT	MEAN	*SD*
must	47	0.28	0.46
certain	56	0.09	0.29
know	52	0.08	0.27

TABLE 2. Data points, mean agreement rate, and standard deviation for each of the three sentences in experiment 1.

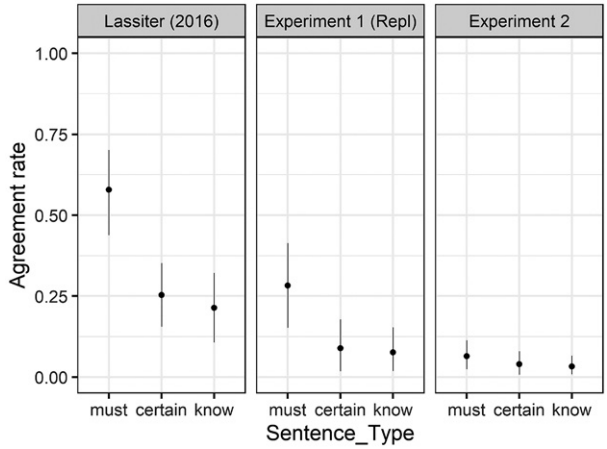

FIGURE 1. Mean ratings in Lassiter 2016, experiment 1, and experiment 2 for *must*, *certain*, and *know*. Error bars indicate bootstrapped 95% confidence intervals.

First, we observed that the proportion of 'agree' responses we obtained was lower in all three conditions than in Lassiter 2016: 0.28 for *must*, 0.09 for *certain*, and 0.08 for *know*; we do not know for sure why that is the case. Perhaps it is because Lassiter did not have a comprehension question in his original design, which might have increased noise (Lassiter, p.c.). Next, we observed that the proportion of agreement for *must* (0.28) was numerically higher than for *certain* (0.09) and for *know* (0.08), and we asked whether this difference was statistically significant. To this end, we analyzed the agreement responses of the three sentence types with a logistic regression model with the factor SENTENCE TYPE as an orthogonal contrast-coded fixed effect (contrast 1: Must = −0.66, Certain = 0.33, Know = 0.33; contrast 2: Must = 0, Certain = −0.5, Know = 0.5) using the R function glm. Participants were more likely to agree with the *must* sentence type than with *certain* and *know* ($\beta = -1.4870$, $z = -3.071$, $p = 0.00214$). Agreement rates did not differ significantly between *certain* and *know* ($\beta = -0.1625$, $z = -0.232$, $p = 0.81649$). Furthermore, we performed pairwise comparisons, using the R package emmeans, showing that there was a significant difference between *must* and *certain* ($\beta = 1.391$, $z = 2.4333$, $p = 0.04$) and *must* and *know* ($\beta = 1.553$, $z = 2.526$, $p = 0.03$). Thus, we successfully replicated Lassiter's (2016) finding that *must* is endorsed significantly more than *certain* and *know*, which do not differ from each other, although the observed rates of agreement were lower overall than in Lassiter 2016. The full output of the statistical analyses can be found in the online supplemental materials.[8]

DISCUSSION. Our replication showed that in a context of almost certainty about *p*: (i) about one third of participants agreed with *must p*, (ii) almost no participant agreed with

[8] The supplemental materials are available at http://muse.jhu.edu/resolve/214.

either *it is certain that p* or *we know that p*, and (iii) the proportion of participants who agreed with *must p* was significantly higher than the proportion who agreed with either *it is certain that p* or *we know that p*. We agree with Lassiter 2016 that these findings do not support the logical *must* hypothesis, but we disagree with the conclusion that they support the probabilistic *must* hypothesis. Specifically, we consider the probabilistic hypothesis at odds with the finding that only one third of participants agreed with *must p*, because this hypothesis predicts an agreement rate very close to 1 based on the assumption that plausibly almost every participant judged *p* to be very likely in the given scenario.

Furthermore, the finding that more people agreed with *must p* than with *it is certain that p* and *we know that p* only suggests that *must p* is weaker than predicted by the logical *must* hypothesis; it does not specifically suggest that the probabilistic *must* hypothesis is the right account for the weakness. In fact, the other two hypotheses reviewed in the introduction are also consistent with these data. According to the hyperbolic logical *must* hypothesis, the minority who agreed with *must* in the context were exaggerating in judging as certain an event that is only probable; in contrast, the majority who did not agree were more careful, in that they recognized that the event of Bill not winning the lottery is not absolutely warranted given the provided information. According to the evidential *must* hypothesis, those who agreed with *must* in the context judged the provided information sufficient to conclude that Bill did not win the lottery, without necessarily judging such a conclusion to be certain, while those who did not agree were more cautious and did not want to jump to such a conclusion.

Next, we aimed to discriminate between the hyperbolic logical hypothesis and the probabilistic hypothesis (we postpone the assessment of the evidential hypothesis to experiment 3). Note that Lassiter's (2016) task design (one sentence per participant without baselines) prompts uncertainty on the part of the reader as to what is intended by 'agreeing' with a statement. In fact, we can think of at least two interpretations of the experimental question 'Do you agree with this sentence in the given context?': some people may interpret it as 'Is this sentence something that one might say in the given context?' (this question would tap the so-called assertability conditions of a sentence), whereas others may interpret it as 'Is this sentence true in the given context?' (which would tap the truth conditions of a sentence). We refer to the first interpretation as the 'assertability task' and to the second as the 'truth-value judgment task'. In general, a positive answer to the assertability task (the sentence might be said in the context) implies a positive answer to the truth-value judgment task (the sentence is true in the context).[9] However, it is possible that participants may consider some types of sentences to be something one might say, yet judge them to be false. For example, this may be the case for sentences expressing a maximal value on a given scale, such as *All my friends hate me*: in informal talk people might say this sentence to rhetorically overstate their feeling of frustration after receiving criticism from a couple of friends, but it is very likely that the speaker themself judges this sentence as not literally true. In this respect, the assertability task is more likely to elicit answers based on this informal way of talking than the truth-value judgment task is, which is better suited for targeting answers based on the literal meanings of sentences. Thus, we speculate that sentences expressing a maximal value on a given scale are more likely to be agreed with

[9] But not vice versa, as shown by the case of sentences containing scalar implicature triggers like *Some of my kids went to college*: one may judge this sentence to be true in a situation where all of the speaker's kids went to college but not as something one might say in the same situation.

under the assertability than under the truth-value judgment interpretation of the task in contexts where a near-maximal value of the scale is defined.

This task feature is crucial for testing the hyperbolic logical hypothesis, which predicts that if participants are induced to focus on the literal meaning of the words contained in the target sentence, then they will converge on the literal strong meaning of *must p* ('it is necessary that *p*'). Consequently, the hyperbolic logical hypothesis predicts that the proportion of 'agree' choices for *must p* in a truth-value judgment task would be lower than in the original experiment (where the task is potentially ambiguous between the assertability task and the truth-value judgment task interpretations) and not different from that of *certain* and *know*. By contrast, the probabilistic hypothesis (*must p* = 'the probability of *p* is very high') predicts no effect induced by disambiguating the task as a truth-value judgment task: if anything, participants are expected to endorse *must p* at a higher rate than *certain p* or *know p* after assessing that it is true that *p* is indeed a highly probable conclusion but not 100% warranted in the given context.

So, the next question is: how do we prompt participants to interpret the task as a truth-value judgment task? This question led us to design experiment 2.

2.2. Experiment 2: including multiple example sentences. In experiment 2, we aimed to assess the hyperbolic logical *must* hypothesis and the probabilistic *must* hypothesis in a task that participants are induced to interpret as a truth-value judgment task. Specifically, we compared Lassiter's original between-subjects design to a within-subjects design, where each participant rated the three critical conditions of experiment 1—'must', 'certain', and 'know'—together with some clearly true and clearly false sentences as baselines: given that participants can only choose between two response options ('agree' or 'disagree'), they would be prompted to assign each of the three target sentences to one of two groups, depending on whether the sentence is like the clearly true sentences (in which case it would be rated with 'agree') or like the clearly false sentences (rated with 'disagree').[10] The probabilistic hypothesis predicts that the agreement rate for *Bill must not have won the raffle* would be like that of the clearly true sentences, while the hyperbolic logical *must* hypothesis predicts that it would be like that of the clearly false sentences. Note that the evidential *must* hypothesis is compatible with either output. Indeed, under this hypothesis, the task would prompt participants to ask themselves whether they would conclude that *p* based on the contextual information, which does not yield a quantifiable prediction: we do not possess a background theory of humans' inferential behavior that would allow us to make a clear prediction about the rate of people who would conclude that Bill did not win the lottery based on the contextual information provided in this experiment.[11] The materials, data, and statistical analyses relevant to this experiment can be found in the OSF project https://osf.io/ukp2w/.

[10] In making this manipulation we were inspired by previous experimental work showing how contextual features of a task affect participants' behavior. For example, studies on scalar implicatures and presuppositions have shown that having participants explicitly evaluate the target implicature/presupposition trigger together with relevant alternatives affects participants' computation rate of the critical implication (e.g. Foppolo et al. 2012, Tonhauser et al. 2013, Skordos & Papafragou 2016, Zehr & Schwarz 2018), whereas other work has shown that participants might be inclined to reject a statement if the experimenter does not recreate the appropriate discourse conditions for its felicitous production (e.g. Syrett 2015, Syrett & Koev 2015, Syrett & Brasoveanu 2019).

[11] We also investigated the effect of negation in two experiments that adopted the same methodology as experiment 1 (sentence manipulated between subjects) and experiment 2 (sentence manipulated within subjects) but without the sentential negation particle in the three target sentences. We found no effect of negation on the agreement responses. The results of these two variants are reported in the supplemental materials as experiment 6 (sentence manipulated between subjects) and experiment 7 (sentence manipulated within subjects).

METHODS.

Participants. We recruited 180 Amazon Mechanical Turkers, dividing them into five groups of thirty-six and assigning them to one of five pseudo-randomized orders (reported in Table 3), created by varying the order of presentation of the three experimental sentences.

ORDER 1	ORDER 2	ORDER 3	ORDER 4	ORDER 5
one	probable	one	certain	probable
two	winner	must	1000	must
know	one	chance	probable	winner
must	certain	know	must	two
probable	chance	probable	know	one
1000	two	two	one	1000
winner	1000	certain	chance	certain
certain	know	1000	winner	know
chance	must	winner	two	chance

TABLE 3. The five pseudo-randomized orders of presentation of the nine sentences in experiment 2.

Materials and design. The story defining the scenario was the same as in Lassiter 2016. The nine sentences that were seen by each participant are listed in 11. Note that we included among the clearly true items the statements *It is highly probable that Bill did not win the raffle* ('probable') and *There is a slight chance that Bill won the raffle* ('chance'): recall that the probabilistic hypothesis predicts that *must p* is truth-conditionally equivalent to *it is highly probable that p* and, consequently, is compatible with *there is a chance that not-p*.

(11) The nine sentences read by participants (within-subjects) in experiment 2
Experimental items
a. Bill must not have won the raffle. (*must*)
b. It is certain that Bill did not win the raffle. (*certain*)
c. We know that Bill did not win the raffle. (*know*)
Clearly true control items
d. It is highly probable that Bill did not win the raffle. (*probable*)
e. There is a slight chance that Bill won the raffle. (*chance*)
f. Bill bought exactly one ticket in the raffle. (*one*)
g. 1000 different people bought one lottery ticket each in the raffle. (*1000*)
Clearly false control items
h. Mary bought two tickets in the raffle. (*two*)
i. The winner will be announced tomorrow. (*winner*)

PREDICTIONS. The dependent measure was the proportion of 'agree' choices for each sentence. The hyperbolic logical *must* hypothesis assumes that *Bill must not have won the raffle* in its literal meaning is truth-conditionally equivalent to *It is certain that Bill did not win the raffle* and *We know that Bill did not win the raffle*, which are expected to be judged as false in the experimental context (the conclusion that Bill did not win the lottery is not warranted given that context). Therefore, the hyperbolic hypothesis predicts that participants will agree with *must* at a rate not significantly different from the rate of *certain* and *know* and numerically close to the expected rate of the clearly false baselines (i.e. very close to zero).

The probabilistic *must* hypothesis assumes that *Bill must not have won the raffle* in its literal meaning is truth-conditionally equivalent to *It is highly probable that Bill did not win the raffle*—which is expected to be judged as true. Therefore, the probabilistic

hypothesis predicts that participants will agree with *must* at a rate significantly higher than the rates for *certain* and *know* and numerically close to the expected rate of the clearly true baselines (i.e. very close to 1).

The evidential *must* hypothesis assumes that *Bill must not have won the raffle* in its literal meaning is truth-conditionally equivalent to *I conclude that Bill did not win the raffle*, whose agreement rate in the experimental context is not predictable. Therefore, the evidential hypothesis does not make any predictions in this experiment and is compatible with any output.

Results. We filtered out results from fifty-five participants because they did not indicate English as their native language or USA as their country, failed to correctly answer the comprehension question, or participated in more than one condition. This left 125 participants. The numbers of data points, mean agreement rates, and standard deviations for the three experimental sentences are reported in Table 4. The mean agreement ratings for the three experimental sentences are plotted in Fig. 1 above (rightmost panel). The mean agreement ratings of all nine sentences are plotted in Figure 2.

SENTENCE TYPE	COUNT	MEAN	*SD*
must	123	0.07	0.25
certain	125	0.04	0.20
know	121	0.03	0.18

Table 4. Data points, mean agreement rate, and standard deviation for each of the three experimental sentences in experiment 2.

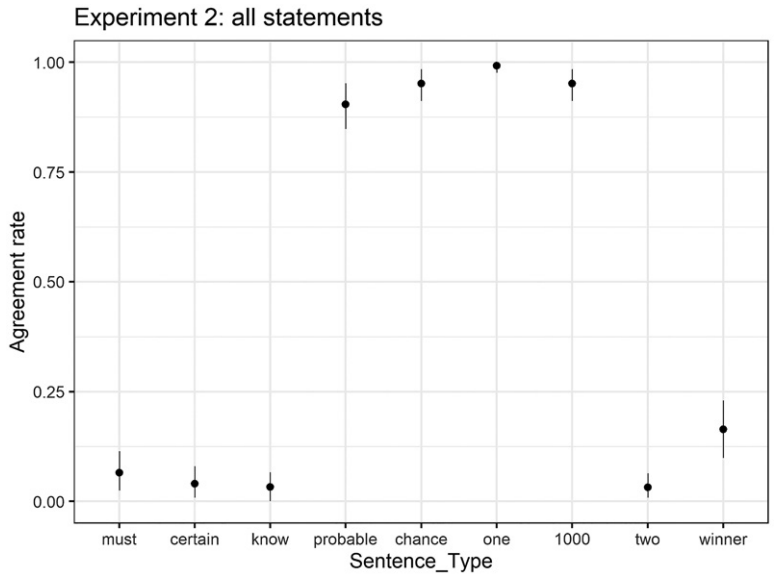

Figure 2. Mean ratings in experiment 2 for all nine sentences listed in 11. Error bars indicate bootstrapped 95% confidence intervals.

First, we observed that the agreement rates of the baselines patterned as expected: the agreement rates of the four clearly true statements (*1000*, *chance*, *one*, *probable*) were close to the ceiling, whereas those of the two clearly false statements (*two*, *winner*) were

close to the floor. Thus, the results of the baseline statements suggest that participants were performing the task by paying attention to the literal meaning of the sentences and that nothing about the task pushed people toward lower agreement across the board. Next, we observed that the mean rating of *must* in experiment 2 (0.07) was numerically lower than in experiment 1 (0.28) and very close to the mean ratings of *certain* and *know* in experiment 2 (0.03 and 0.04, respectively). So, we asked two questions: (i) whether the mean rating of *must* differed from the mean ratings of *certain* and *know* in experiment 2, and (ii) whether the probability of obtaining the observed rating decrease for *must* from experiment 1 to experiment 2 was different from chance.

To answer question (i) we analyzed the agreement responses of the three critical sentences in experiment 2 with a logistic regression model with the factor SENTENCE TYPE as an orthogonal contrast-coded fixed effect (contrast 1: Must = −0.66, Certain = 0.33, Know = 0.33; contrast 2: Must = 0, Certain = −0.5, Know = 0.5). Agreement rates did not differ between *must* and *certain* and *know* ($\beta = -0.6177$, $z = -1.222$, $p = 0.222$). To answer question (ii), we analyzed the agreement responses of the three critical sentences in experiment 1 and experiment 2 combined with a logistic regression model with the factor Sentence type as an orthogonal contrast-coded fixed effect and the factor EXPERIMENT as an effects-coded fixed effect (Experiment 1 = −0.5; Experiment 2 = 0.5). There was a main effect of Experiment such that the rates of agreement were significantly lower in experiment 2 than in experiment 1 ($\beta = -1.160$, $z = -3.180$, $p < 0.01$); more specifically, after conducting pairwise comparisons, we found that the rates of *must* endorsement were lower in experiment 2 than in experiment 1 ($\beta = 1.734$, $z = 3.533$, $p < 0.001$), suggesting that the within-subjects presentation of sentences reduced participants' agreement with the *must* statement.[12]

Following the suggestion of a referee, we also asked whether the order of presentation of the nine sentences had an effect on the rate at which participants agreed with the *must* sentence. To answer this question, we conducted an analysis with order 1–5 as the only predictor of the *must* endorsement rate. A χ^2 test comparing the residual deviance to null deviance reveals a marginally significant effect of presentation order ($p = 0.056$). In order not to overinterpret a marginally significant result, we did not explore these differences further. See the supplemental materials for the full output of the statistical analyses.

DISCUSSION. In experiment 2, we changed Lassiter's original one-sentence task to a multiple-sentence task in which each participant judged the three experimental statements (*Bill must not have won the raffle*, *It is certain that Bill did not win the raffle*, and *We know that Bill did not win the raffle*) while they were simultaneously displayed on the screen, together with some clearly true—including *It is highly probable that Bill did not win the raffle*—and clearly false baseline statements. By making these changes, we aimed to prompt participants to assign sentences to two groups: an 'agree' group, including sentences like the clearly true baselines, and a 'disagree' group, including sentences like the clearly false baselines. We found that the agreement rate for *must* did not significantly differ from the agreement rates for *certain* and *know* and patterned with the agreement rates of clearly false sentences, whereas *probable* and *chance* patterned with clearly true statements, as expected.

[12] Following the suggestion of a referee, we also analyzed the *must* data independently with Experiment as an effect-coded fixed effect, and we found again that the rates of *must* endorsement were lower in experiment 2 than in experiment 1 ($\beta = -1.733$, $z = -3.533$, $p < 0.001$).

We take these findings to suggest that, for English speakers, judging an event to be highly probable in a given scenario is not a sufficient condition for endorsing *must p* in that scenario, hence falsifying the probabilistic *must* hypothesis. However, these findings do not discriminate between the hyperbolic logical *must* and the evidential hypotheses. The hyperbolic logical hypothesis would explain the results by assuming that participants converged on the literal meaning of *must p* (*p* is a necessary event) and, consequently, converged on rating *Bill must not have won the raffle* as false based on the contextual information (it is false that it is certain that Bill did not win the lottery, given the contextual information). The evidential hypothesis would explain the results by assuming that participants converged on judging the contextual information as insufficient to conclude that Bill did not win the lottery. Therefore, we take our findings from experiment 2 to be compatible with both the hyperbolic logical *must* hypothesis and the evidential *must* hypothesis.

So, how do we discriminate between these two hypotheses? We started by examining the nature of the scenario designed by Lassiter (2016), which we kept constant across experiments 1 and 2. Recall that the story was designed such that the probability of the event of Bill having won the lottery is known and very small (one chance out of 1,000), based on which one may be induced to conclude that he did not win the lottery. But what type of inference would one be performing in this case? This is an instance of what philosophers call 'inductive inference', namely a nonlogically valid inferential pattern 'based purely on statistical data, such as observed frequencies of occurrences of a particular feature in a given population' (Douven 2017). A common practice in the philosophical literature is to distinguish within the class of nonlogically valid inferences the inductive type from the abductive type (also known as 'inference to the best explanation'; cf. Harman 1965): 'in abduction there is an implicit or explicit appeal to explanatory considerations, whereas in induction there is not; in induction, there is *only* an appeal to observed frequencies or statistics' (Douven 2017). A good example of abductive inference is the following: 'You happen to know that Tim and Harry have recently had a terrible row that ended their friendship. Now someone tells you that she just saw Tim and Harry jogging together. The best explanation for this that you can think of is that they made up. You conclude that they are friends again' (Douven 2017). Thus, abductive conclusions and inductive conclusions are alike in being obtained through nonlogically valid inferential strategies, but they crucially differ in whether the conclusion is TRIGGERED BY THE NEED TO EXPLAIN some other known fact.

Reexamining the results of experiment 2 in these terms, we see that English speakers do not agree with *must p* in a context where *p* is the conclusion of an inductive inference. We designed experiment 3 in order to determine whether speakers also converge on not believing that *p* if *p* is the conclusion of an inference to the best explanation—an abductive inference.

2.3. EXPERIMENT 3: BEST EXPLANATIONS VS. INDUCTIVE CONCLUSIONS. In experiment 3, we aimed to assess two potential accounts of the experiment 2 finding that participants overwhelmingly did not endorse *must p* in a context where *p* describes an event whose occurrence is very likely based on a pure statistical computation: according to the hyperbolic logical *must* hypothesis, participants behaved so because such an event is not certain, whereas according to the evidential hypothesis, they behaved so because statistically strong support for an event is not enough to make them conclude (i.e. form the belief) that such an event happened.

How would participants behave if the same event, with the same degree of statistical support, were presented to them as also being the most plausible explanation for some other event? As mentioned in the introduction, numerous findings in the psychology of reasoning suggest that people are sensitive to an EXPLANATION EFFECT when estimating the likelihood of an event: the higher the quality (the simplicity, the breadth, and the coherence) of an explanation, the higher people will estimate its likelihood. We therefore expect participants to endorse a statement like *I conclude that Bill did not win the raffle* more in a context where the event of Bill not winning the lottery represents the best explanation for some other fact than in a context where the same event is just a probabilistically supported conclusion. And they will do so even if they judge the event as not certain in the explanatory context.

We take the two hypotheses—hyperbolic logical *must* and evidential *must*—as differing in terms of their predictions in a task comparing inductive-type and explanatory-type contexts. The hyperbolic logical *must* hypothesis predicts that even in the explanatory context participants will agree with *Bill must not have won the raffle* as much as with *It is certain that Bill did not win the raffle*, regardless of whether their agreement rate with *I conclude that Bill did not win the raffle* is higher: in the explanatory as well as in the inductive context, that Bill did not win the lottery is an uncertain event. Instead, the evidential hypothesis predicts that participants' endorsement rate of *Bill must not have won the raffle* will increase in the explanatory context and align with the agreement rate of *I conclude that Bill did not win the raffle*. For completeness, we also tested whether across the two contexts the endorsement rate of *Bill must not have won the raffle* aligns with that of *It is highly probable that Bill did not win the raffle*, as the probabilistic hypothesis would predict.[13] The materials, data, and statistical analyses relevant to this experiment can be found in the OSF project https://osf.io/ukp2w/.

METHODS.

Participants. We recruited 140 participants on Prolific, aiming to have at least 120 good participants (assuming that ~10% might make an error on one of the four sanity-check sentences). We divided participants into two groups of seventy and assigned them to one of the two context conditions. Again, as in the previous two experiments, the dependent measure was the proportion of 'agree' choices for each sentence. We excluded eight participants from analysis because they did not rate all of the baselines as expected, which left us with 132 participants (sixty-seven in the inductive condition and sixty-five in the explanatory condition).

Materials and design. We adopted a 2 × 4 mixed design by crossing the factor CONTEXT (two levels: Inductive, Explanatory; see 12) manipulated between subjects and the factor SENTENCE TYPE (eight levels: Must, Conclude, Certain, Probable, One, X1000, Two, Winner; see 13) manipulated within subjects, as in experiment 2. First, each participant saw either Lassiter's original context, defining 'Bill did not win the raffle' as a highly plausible but not certain conclusion based on mere probabilistic reasoning (see 12a), or a context defining the same conclusion as a very good explanation for a

[13] We also ran another variant of this experiment in which we crossed the two contexts with the number of sentences rated by participants (one-sentence task vs. multiple-sentence task). The results of the two multiple-sentence conditions of this variant are similar to those of experiment 3: crucially, the agreement rate of *must* was higher in the explanatory than in the inductive condition and was higher than the agreement rates of *certain* or *know* in the explanatory condition. The same pattern was observed in the one-sentence conditions as well. The full results are reported in the supplemental materials as experiment 8.

fact described in the story, but still not certain (see 12b). Next, each participant saw all eight sentences in 13 on the screen simultaneously in a random order and chose between the response options 'agree' and 'disagree' for each. The critical sentences are *Bill must not have won the raffle*, *It is certain that Bill did not win the raffle*, *It is highly probable that Bill did not win the raffle*, and *I conclude that Bill did not win the raffle* (which would allow us to keep track of participants' willingness to form the belief that *p* based on the context). The clearly true and false baselines are there to provide sanity checks: we analyzed only the data from participants who rated all four clearly true and clearly false baselines as expected.

(12) The two stories read by participants (between-subjects) in experiment 3
   a. Lassiter's 2016 context (INDUCTIVE) (suggesting a conclusion supported by a probabilistic computation)
   Yesterday, Bill bought a single ticket in a raffle with 1000 total tickets. There were also 999 other people who bought one ticket each. That is, the tickets were distributed like this: People holding one ticket: Bill, Mary, Jane, … [997 more]. The drawing was held last night, and the winner will be announced this evening.
   b. Modified version (EXPLANATORY) (suggesting a conclusion that is simple, coherent, and consistent with prior biases)
   Yesterday, Bill bought a single ticket in a raffle with 1000 total tickets. There were also 999 other people who bought one ticket each. That is, the tickets were distributed like this: People holding one ticket: Bill, Mary, Jane, … [997 more]. The drawing was held last night. **Today, you meet Bill and he looks a little bit disappointed**.

(13) The eight sentences read by participants (within-subjects) in experiment 3
   Experimental items
   a. Bill must not have won the raffle. (*must*)
   b. I conclude that Bill did not win the raffle. (*conclude*)
   c. It is certain that Bill did not win the raffle. (*certain*)
   d. It is highly probable that Bill did not win the raffle. (*probable*)
   Clearly true baselines
   e. Bill bought exactly one ticket in the raffle. (*one*)
   f. 1000 different people bought one lottery ticket each in the raffle. (*1000*)
   Clearly false baselines
   g. Mary bought two tickets in the raffle. (*two*)
   h. The winner will be announced tomorrow. (*winner*)

PREDICTIONS. The dependent measure was the proportion of 'agree' choices for each sentence. We expected that the endorsement rate of *certain* would be close to floor and that of *probable* would be close to ceiling in both contexts (based on the findings of experiment 2 for the inductive context and on our intuition for the explanatory context). We further expected the endorsement rate of *conclude* to be much higher in the explanatory than in the inductive condition and much higher than that of *certain* in the explanatory condition (because of the existence of an explanation effect). Crucially, the three hypotheses make the following predictions about the agreement rate of *must*:

- The hyperbolic logical *must* hypothesis predicts that *must* will pattern with *certain* in both the inductive and the explanatory contexts.

- The probabilistic *must* hypothesis predicts that *must* will pattern with *probable* in both the inductive and the explanatory contexts.
- The evidential *must* hypothesis predicts that *must* will pattern with *conclude* in both the inductive and the explanatory contexts: that is, *must* will be higher in the explanatory condition than in the inductive condition, and *must* will be higher than *certain* in the explanatory condition.

RESULTS. The mean agreement rates for the critical sentences are reported in Table 5 and plotted in Figure 3.

INFERENCE TYPE	SENTENCE TYPE	*N*	MEAN	*SD*
INDUCTIVE	must	67	0.21	0.41
	conclude	67	0.37	0.49
	certain	67	0.09	0.29
	probable	67	0.97	0.17
EXPLANATORY	must	65	0.82	0.39
	conclude	65	0.89	0.31
	certain	65	0.40	0.49
	probable	65	0.97	0.17

TABLE 5. Data points, mean agreement rate, and standard deviation for each of the four experimental sentences in experiment 3.

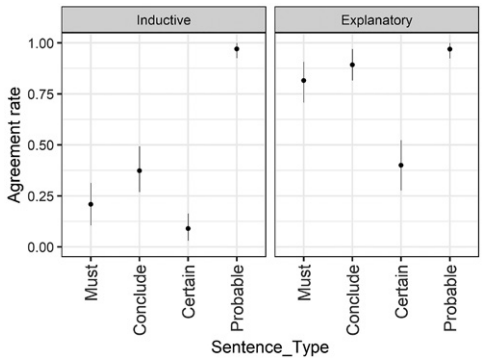

FIGURE 3. Mean ratings from participants who correctly rated the baselines in experiment 3 for *must*, *conclude*, *certain*, and *probable*. Error bars indicate bootstrapped 95% confidence intervals.

We analyzed the agreement responses for the four critical sentences with a logistic mixed-effects regression model (using the glmer function from the lme4 package). The model included the factor Sentence type (Must, Conclude, Certain, Probable) as a dummy-coded predictor (with Must as the reference level for one model, and Conclude as the reference level in a follow-up analysis), the factor Context as an effects-coded predictor (Inductive = −0.5; Explanatory = 0.5), their interaction, and random intercepts for participants. We followed up this analysis with pairwise comparisons of the four sentences across the two contexts using the R function emmeans.

First, we asked whether our manipulation of the contextual information was successful. Indeed, we found a main effect of context type such that, overall, the agreement rates of the sentences were significantly higher in the explanatory context than in the inductive context ($\beta = 6.60$, $z = 3.86$, $p < 0.001$), which suggests that the explanatory

context induced higher agreement rates overall, supporting the existence of an explanation effect.

Next, we asked whether our expectations about the agreement rates of *probable*, *certain*, and *conclude* were confirmed. Concerning *probable*, we observed that its agreement rate was as expected, close to ceiling across the two contexts (0.97 in both). Concerning *certain*, while in the inductive condition its agreement rate was close to floor, as expected (0.09), in the explanatory condition it was higher than expected (0.41), which may be due to the explanation effect.

Concerning *conclude*, we observed that its agreement rate was much higher in the explanatory (0.89) than in the inductive (0.37) condition (pairwise comparison Conclude inductive vs. Conclude explanatory: $\beta = -6.20$, $z = -3.58$, $p < 0.001$), which suggests that the 'best explanation' nature of the event of Bill not winning the lottery prompted more participants to conclude that such an event happened than its objectively high probability alone did. In both the inductive and the explanatory conditions, the rate of agreement with *conclude* differed from both *certain* (Inductive: $\beta = -3.59$, $z = -3.94$, $p < 0.001$; Explanatory: $\beta = -5.79$, $z = -3.98$, $p < 0.001$) and *probable* (Inductive: $\beta = -7.19$, $z = -5.08$, $p < 0.001$; Explanatory: $\beta = -2.62$, $z = -2.06$, $p < 0.05$). In the inductive condition, the agreement rate for *conclude* (0.37) was closer to that of *certain* (0.09) than to that of *probable* (0.97), whereas in the explanatory condition this pattern was reversed, with the agreement rate of *conclude* (0.89) closer to that of *probable* (0.97) than to that of *certain* (0.40). We found an interaction effect between Sentence type–Contrast 3 (Conclude vs. Probable) and the context condition (Inductive vs. Explanatory), such that participants agreed with *probable* more than with *conclude* more so in the inductive than in the explanatory context ($\beta = -4.59$, $z = -2.85$, $p < 0.01$). However, the interaction between Sentence type–Contrast 2 (Conclude vs. Certain) and the two contexts did not reach significance ($\beta = -2.20$, $z = -1.68$, $p = 0.09$).

Lastly, we investigated the agreement rates of *must* to assess the predictions of the three hypotheses. Concerning *must* on its own, we observed that its agreement rate was much higher in the explanatory (0.82) than in the inductive (0.21) condition (pairwise comparison Must inductive vs. Must explanatory: $\beta = -6.59$, $z = -3.85$, $p < 0.001$), which suggests that people's endorsement of *must p*—like that of *certain p* and *I conclude that p*—is sensitive to the explanation effect. Concerning *must* in relation to the other three sentences, first we found that the agreement rate of *must* was different from that of each of the other three sentences on average across the two contexts: higher than that of *certain* ($\beta = -3.24$, $z = -3.96$, $p < 0.001$), lower than that of *probable* ($\beta = 6.36$, $z = 5.11$, $p < 0.001$), and lower than that of *conclude* ($\beta = 1.46$, $z = 2.81$, $p < 0.01$). However, the difference in rates between *must* and *conclude* was much smaller than the difference in rates between *must* and *certain* or *probable*.

We also found an interaction effect between Sentence type–Contrast 1 (Must vs. Certain) and the context condition (Inductive vs. Explanatory) such that participants agreed with *certain* less than with *must* more so in the explanatory than in the inductive context ($\beta = -2.60$, $z = -2.09$, $p < 0.05$). Similarly, Sentence type–Contrast 2 (Must vs. Probable) interacted significantly with Context such that participants agreed with *probable* more than with *must* more so in the inductive than in the explanatory context ($\beta = -4.97$, $z = -3.01$, $p < 0.01$). No interaction was detected between Sentence type–Contrast 3 (Must vs. Conclude) and Context ($\beta = -0.40$, $z = -0.42$, $p = 0.67$). Indeed, the estimated effect sizes for how much participants agreed with *must* relative to *conclude*

were very similar in the inductive ($\beta = -1.66$) and explanatory ($\beta = -1.26$) conditions. Similarly to *conclude*, in the inductive condition, the agreement rate of *must* was much closer to that of *certain* (pairwise comparison Must inductive vs. Certain inductive: $\beta = 1.94$, $z = 2.43$, $p < 0.05$) than to that of *probable* (pairwise comparison Must inductive vs. Probable inductive: $\beta = -8.84$, $z = -5.46$, $p < 0.001$), whereas in the explanatory condition this pattern was reversed, with the agreement rate of *must* much closer to that of *probable* (pairwise comparison Must explanatory vs. Probable explanatory: $\beta = -3.88$, $z = -2.85$, $p < 0.01$) than to that of *certain* (pairwise comparison Must explanatory vs. Certain explanatory: $\beta = 4.53$, $z = 3.73$, $p < 0.001$). These findings suggest that participants' agreement rates for *must* and *conclude* aligned in the way they patterned relative to the rates for *certain* and *probable* across the two contexts.

Overall, these findings suggest that, although the agreement rate of *must* is statistically lower than that of *conclude*, the two agreement rates pattern together in both contexts and differ similarly from those of *certain* and *probable* across the two contexts.

Discussion. In experiment 3, participants were asked to decide whether they agree with each of the four sentences *Bill must not have won the raffle* ('must'), *I conclude that Bill did not win the raffle* ('conclude'), *It is certain that Bill did not win the raffle* ('certain'), and *It is highly probable that Bill did not win the raffle* ('probable'), plus two clearly true and two clearly false baselines, given one of these two contexts: either Lassiter's original context defining the proposition 'Bill did not win the raffle' as a nonlogical conclusion derived through a probabilistic computation (inductive context) or a context defining the same proposition as a nonlogical conclusion but a plausible explanation for some of the described facts (explanatory context).

We found that participants agreed with *conclude* at a much higher rate in the explanatory than in the inductive context, suggesting that a highly probable event that also explains well some other event is believed more than an event that is highly probable just based on a statistical estimate. The endorsement rate of *must* was: (i) much lower than that of *probable* and close to that of *certain* in the inductive context, (ii) much higher than that of *certain* and closer to that of *probable* in the explanatory context, and (iii) similar to (although slightly lower than) that of *conclude* in both contexts. We take these findings to suggest that comprehenders' endorsement of *must p* in a context is more affected by whether they would say they conclude that *p* in that context than whether they would say that *p* is certain or probable in that context. More specifically, these findings suggest that speakers say *must p* as well as *I conclude that p* not whenever they judge *p* as objectively highly probable and not only when they judge *p* as objectively certain but whenever their subjective probability of *p* passes the threshold above which they would adopt *p* as one of their beliefs. Our findings show that one of the factors that play a role in the computation of such a subjective probability threshold of an event is whether, in addition to being objectively highly probable, this event is also the most plausible explanation for another event whose occurrence would otherwise be hard to motivate. Overall, these findings support the evidential account of *must p* over the hyperbolic logical account and the probabilistic account.

Interestingly, we found that the endorsement rate of *certain* was also higher in the explanatory than in the inductive condition, suggesting that even the computation of the objective certainty of an event is sensitive to an explanation effect: for some people, a conclusion that represents a highly probable, good, and simple explanation of some other facts is certain.

So far, we have considered only the English *must*, which raises the question of how generalizable our findings are to other languages: is a general property of necessity auxiliary verbs that they are used as inferential evidentials? To start answering this question, we attempted to reproduce the findings of experiment 3 in an experiment with Italian speakers featuring the epistemic use of *deve p*, which we describe in the following section.

2.4. Experiment 4: best explanations vs. inductive conclusions in italian. In experiment 4 we attempted to reproduce in Italian the findings of experiment 3. In experiment 3 we compared the endorsement rate of *must p* to that of *I conclude that p*, which closely approximates the meaning of *must p* under the evidential hypothesis. But we could not compare epistemic *must* to any grammatical inferential because there is no such type of expression in English. In Italian, there exists a morpheme that has been argued to behave like an inferential evidential, namely the future morphology in its nontemporal uses (cf. Squartini 2001, Pietrandrea 2005, Eckardt & Beltrama 2019, Frana & Menéndez-Benito 2019). A naturally occurring example of an Italian utterance where the future is interpreted as an inferential is given in 14.

(14) The Italian future in its reading as an inferential
Del resto, La Regressione non **sarà** certo stato scelto a caso, come sottotitolo.[14]
'After all, "La Regressione" will not have been chosen accidentally as a subtitle.'

In Italian we can therefore compare the endorsement rate of a sentence containing the verb *dovere* (the Italian counterpart of *must*) in its epistemic reading to a sentence containing a grammatical inferential across the inductive and explanatory contexts. The materials, data, and statistical analyses relevant to this experiment can be found in the OSF project https://osf.io/ukp2w/.

Methods.

Participants. We recruited 140 Italian native speakers on Prolific, aiming to have 120 good participants (assuming that ~10% might make an error on one of the four sanity-check sentences). We excluded twenty-seven participants from analysis because they did not rate all of the baselines as expected, which left us with 113 (fifty-four for the inductive condition and fifty-nine for the explanatory condition). We divided participants into two groups of seventy and assigned them to one of the two context conditions.

Materials and design. We translated the two contexts of experiment 3 from English into Italian and changed the name of the protagonist from Bill to Gianni. Thus, in this experiment the critical conclusion suggested by both contexts is *Gianni non ha vinto la lotteria* 'John has not won the raffle'. The stories are given in 15.

(15) The two stories read by participants (between-subjects) in experiment 4
a. Lassiter's 2016 context (suggesting an inductive conclusion)
Ieri Gianni ha comprato un biglietto di una lotteria comprendente 1000 biglietti in tutto. Altre 999 persone hanno comprato un biglietto ciascuna. Quindi, i biglietti sono distribuiti come segue: le persone con un biglietto sono: Gianni, Maria, Sandra, … [altri 997]. L'estrazione dei biglietti è stata effettuata ieri e il vincitore verrà annunciato stasera.

[14] Source: http://www.tvblog.it/categoria/le-calde-notti-di-tvblog

b. Modified version (suggesting an EXPLANATORY conclusion that is simple, coherent, and consistent with prior biases)
Ieri Gianni ha comprato un biglietto di una lotteria comprendente 1000 biglietti in tutto. Altre 999 persone hanno comprato un biglietto ciascuna. Quindi, i biglietti sono distribuiti come segue: le persone con un biglietto sono: Gianni, Maria, Sandra, … [altri 997]. Il vincitore è stato annunciato ieri. Oggi, ti capita di incontrare Gianni che sembra deluso.

We also adopted the sentences of experiment 3 and translated them into Italian (see 16). Crucially, we added the critical sentence 'futuro' containing the inferential future morpheme.

(16) The nine sentences read by participants (within-subjects) in experiment 4
Experimental items
a. Gianni non deve aver vinto la lotteria. (*deve*)
'John must not have won the raffle.'
b. Gianni non avrà vinto la lotteria. (*futuro*)
'John will have not won the raffle.'
c. Deduco che Gianni non ha vinto la lotteria. (*deduco*)
'I deduce that John has not won the raffle.'[15]
d. È certo che Gianni non ha vinto la lotteria. (*certo*)
'It is certain that John has not won the raffle.'
e. È altamente probabile che Gianni non ha vinto la lotteria. (*probabile*)
'It is highly probable that John has not won the raffle.'
Clearly true baselines
f. Gianni ha comprato esattamente un biglietto della lotteria. ('one')
'John has bought exactly one ticket in the raffle.'
g. 1000 persone hanno comprato ciascuna un biglietto della lotteria. ('X1000')
'1000 people have bought one lottery ticket each in the raffle.'
Clearly false baselines
h. Maria ha comprato due biglietti della lotteria. ('two')
'Mary has bought two tickets in the raffle.'
i. Il vincitore verrà annunciato la prossima settimana. ('winner')
'The winner will be announced next week.'

As in experiment 3, we crossed the factor CONTEXT (two levels: Inductive, Explanatory) with the factor SENTENCE TYPE (nine levels: Deve, Futuro, Deduco, Certo, Probable, One, X1000, Two, Winner) in a mixed design, with Context manipulated between subjects and Sentence manipulated within subjects. Each participant saw either the inductive or the explanatory context, with all nine sentences presented simultaneously on the screen in a random order, and they chose between the response options 'agree' and 'disagree' for each sentence. Responses to *deve*, *futuro*, *deduco*, *certo*, and *probabile* are critical, and the clearly true and false baselines are there to provide sanity checks: we analyzed only the data from participants who rated as expected all four clearly true and clearly false baselines.

PREDICTIONS. The dependent measure was the proportion of 'agree' choices for each sentence. Based on the findings of experiment 3, we expected that: (i) the endorsement

[15] We translated the expression *I conclude* from the original English sentence with the verb *deduco* 'I deduce' instead of its Italian cognate *concludo* because the first author as an Italian native speaker judged a sentence with *dedurre* 'deduce' to be more natural than one with *concludere* 'conclude'.

rate of *probabile* would be close to ceiling in both contexts, (ii) the endorsement rate of *certo* would be close to floor in the inductive context and higher in the exploratory context, and (iii) the endorsement rate of *deduco* would be much higher in the explanatory than in the inductive condition and much higher than that of *certain* in the explanatory condition. We also expected the endorsement rate of *futuro* to be similar to that of *deduco*. Crucially, the three hypotheses make the following predictions:

- The hyperbolic logical *must* hypothesis predicts that the endorsement rate of *deve* will be similar to that of *certo* in both the inductive and the explanatory contexts.
- The probabilistic *must* hypothesis predicts that the endorsement rate of *deve* will be similar to that of *probabile* in both the inductive and the explanatory contexts.
- The evidential *must* hypothesis predicts that the endorsement rate of *deve* will be similar to that of *deduco* and *futuro* in both the inductive and the explanatory contexts.

RESULTS. The mean agreement rates and standard deviations for the five experimental sentences are reported in Table 6 and plotted in Figure 4.

INFERENCE TYPE	SENTENCE TYPE	*N*	MEAN	*SD*
INDUCTIVE	deve	54	0.20	0.41
	futuro	54	0.22	0.42
	deduco	54	0.17	0.38
	certo	54	0.00	0.00
	probabile	54	0.83	0.38
EXPLANATORY	deve	59	0.83	0.38
	futuro	59	0.90	0.30
	deduco	59	0.86	0.35
	certo	59	0.29	0.46
	probabile	59	0.98	0.13

TABLE 6. Data points, mean agreement rate, and standard deviation for each of the five experimental sentences in experiment 4.

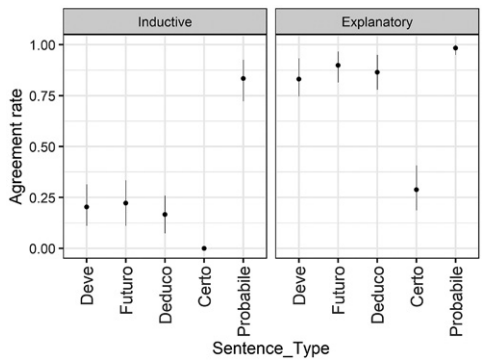

FIGURE 4. Mean ratings from participants who correctly rated the baselines in experiment 4 for *deve*, *futuro*, *deduco*, *certo*, and *probabile*. Error bars indicate bootstrapped 95% confidence intervals.

We could not analyze the agreement responses of the five critical sentences with a logistic mixed-effects regression model (as we did in experiment 3) because the agreement rate in the condition Inductive–Certo was exactly zero (i.e. had 0 variance). So, we first describe the relevant observed patterns in the raw ratings and then report the output

of three analyses, whose combination provides statistical support for those observed patterns.

First, we observed that our expectations about the agreement rates of *probabile*, *certo*, and *deduco* were confirmed. Concerning *probabile*, we observed that its agreement rate was high in both contexts (Inductive: 0.83; Explanatory: 0.98). Concerning *certo*, its agreement rate was at floor (0) in the inductive condition and higher in the explanatory condition (0.29). For *deduco*, we observed that indeed its agreement rate was much higher in the explanatory (0.86) than in the inductive (0.17) condition and much higher than that of *certo* in the explanatory condition.

Next, we observed that our expectation about the agreement rates of *futuro* was also confirmed: its agreement rate was very similar to that of *deduco* in both the inductive condition (Deduco: 0.17; Futuro: 0.22) and the explanatory condition (Deduco: 0.86; Futuro: 0.90).

Lastly, we observed that the agreement rate for *deve* was: (i) different from that of *certo* in the explanatory condition (Deve: 0.83; Certo: 0.29), contrary to what would be predicted by the hyperbolic hypothesis; (ii) different from that of *probabile* in the inductive condition (Deve: 0.20; Probabile: 0.83), contrary to what would be predicted by the probabilistic hypothesis; and (iii) very similar to that of *futuro* and *deduco* in both the inductive condition (Deduco: 0.17; Futuro: 0.22; Deve: 0.20) and the explanatory condition (Deduco: 0.86; Futuro: 0.90; Deve: 0.83), as predicted by the evidential hypothesis.

To offer statistical support for these observations, we ran the following analyses. First, we analyzed the agreement responses from the four critical sentences *deve*, *futuro*, *deduco*, and *probabile* with a logistic mixed-effects regression model with the factor Sentence type defined as a dummy-coded predictor (with Deve as the reference level), the factor Context as an effects-coded predictor (Inductive = −0.5; Explanatory = 0.5), their interaction, and random intercepts for participants. We followed up this analysis with pairwise comparisons of the four sentences across the two contexts using the R function emmeans. Crucially, we found that on average across the two contexts the agreement rate of *deve* was different from (lower than) that of *probabile* ($\beta = 10.50$, $z = 7.09$, $p = 0.09$), but not from that of *futuro* ($\beta = 1.20$, $z = 1.67$, $p < 0.001$) and *deduco* ($\beta = -0.08$, $z = -0.11$, $p = 0.91$). Sentence type–Contrast 3 (Deve vs. Probabile) interacted significantly with Context such that participants agreed with *probabile* more than with *deve* more so in the inductive than in the explanatory context ($\beta = -9.20$, $z = -3.53$, $p < 0.001$). Instead, we found no interaction between Context and Sentence type–Contrast 1 (Deve vs. Futuro) ($\beta = 1.45$, $z = 1.02$, $p = 0.31$) or Sentence type–Contrast 2 (Deve vs. Deduco) ($\beta = 2.01$, $z = 1.39$, $p = 0.17$). For other results see the supplemental materials.

Next, to analyze the relation between *deve* and *certo* in the explanatory condition, we analyzed with a logistic regression model the agreement responses from all five critical sentences, but only from the explanatory context. Crucially, we found that in the explanatory condition the agreement rate of *deve* (0.83) was higher than that of *certo* (0.29; $\beta = -5.96$, $z = -2.09$, $p < 0.05$).

Lastly, we ran a binomial test using the R function binom.test to compare the agreement rate of *deve* to the zero-agreement rate of *certo* in the inductive context. We found that the agreement rate of 0.20 for *deve* given a sample size of fifty-four was significantly different from an agreement rate of 0 ($p < 0.001$).

To summarize, these findings suggest that Italian comprehenders endorse epistemic *deve p* as well as inferential *p*-future in a given context not based on whether they judge

p as certain or probable in the context but based on whether they would conclude (deduce) that *p* in the context.

Discussion. In experiment 4, we attempted to replicate the findings of experiment 3 (which was performed with English participants) with Italian participants. Participants were asked to decide whether they agree with each of five sentences: *Gianni non deve aver vinto la lotteria* 'John must not have won the raffle' ('deve'), *Gianni non avrà vinto la lotteria* 'John will have not won the raffle' ('futuro'), *Deduco che Gianni non ha vinto la lotteria* 'I deduce that John has not won the raffle' ('deduco'), *È certo che Gianni non ha vinto la lotteria* 'It is certain that John has not won the raffle' ('certo'), and *È altamente probabile che Gianni non ha vinto la lotteria* 'It is highly probable that John has not won the raffle' ('probabile'), plus two true baselines and two false baselines. There were two contexts: either Lassiter's original context defining the proposition *Gianni non ha vinto la lotteria* 'John has not won the raffle' as a nonlogical conclusion derived through a probabilistic computation (inductive context), or a context defining the same proposition as a plausible explanation for some of the described facts (explanatory context).

We found that the endorsement rate of *deve* was: (i) much lower than that of *probabile* in the inductive context, (ii) much higher than that of *certo* in the explanatory context, and (iii) similar to that of *futuro* and *deduco* in both contexts. These findings suggest that Italian comprehenders endorse *deve p* as well as inferential *p*-future not whenever they judge *p* as objectively highly probable and not only when they judge *p* as objectively certain but whenever their subjective probability of *p* passes the threshold above which they would conclude (i.e. form the belief) that *p*. Overall, we take these findings to support the evidential account of *deve p* over the hyperbolic logical account and the probabilistic account.

3. General discussion. In this article, we aimed to offer an assessment of the debate about the meaning of necessity auxiliary verbs in their epistemic interpretation, focusing on English *must* and Italian *deve*. We have argued that the proposals advanced in the literature can be grouped into three main positions: the hyperbolic logical hypothesis, the probabilistic hypothesis, and the evidential hypothesis. We reported four comprehension tasks (three with English speakers and one with Italian speakers) asking participants to decide whether they endorse a statement based on the given contextual information. Overall, these experiments suggest the following four conclusions.

A. The endorsement rate of epistemic *must p* in a given context is lower when participants evaluate it together with other statements about *p*'s likelihood and baselines exemplifying the task as a truth-value judgment task than when they evaluate it alone (experiments 1–2).
B. English speakers' endorsement of *must p* and Italian speakers' endorsement of *deve p* (and sentences with the inferential future) in a given context depend on whether such a context leads them to conclude that the event described by *p* happened (experiments 3–4).
C. People conclude that an event happened independently from their estimation of the objective probability of the event: they do not need to be certain about an event to conclude that it happened, and they do not always conclude that an event happened when they judge it to be highly probable (experiments 3–4).
D. People are more often induced to conclude that an event happened if this event is a plausible explanation for another event than if the event is just supported by statistical reasoning (experiments 3–4).

Overall, we take these findings as supporting the evidential *must* hypothesis, which holds that people understand (and use) epistemic *must* and *deve* as inferential evidentials, over the two other alternative hypotheses.

In addition, we take finding A—the agreement rate of epistemic *must p* is lower in the multiple-sentence task than in the one-sentence task—as contributing to the recent literature assessing basic features of sentence judgment tasks adopted in syntactic and semantic research (e.g. Katsos & Bishop 2011, Foppolo et al. 2012, Gibson & Fedorenko 2013, Tonhauser et al. 2013, Syrett 2015, Syrett & Koev 2015, Skordos & Papafragou 2016, Sprouse & Almeida 2017, Zehr & Schwarz 2018, Jasbi et al. 2019, Syrett & Brasoveanu 2019, Davidson 2020, Marty et al. 2020, Waldon & Degen 2020). We assume that when a statement is evaluated alone, there is uncertainty among participants as to what the task might be. One possibility is a truth-value task, but another interpretation is that it might be an assertability task (e.g. 'Is it possible that someone might say this sentence in this context?'), which might have higher agreement. But when clearly true/false statements are provided as baselines, participants infer that the task is not the assertability task, because no one would say the false baselines, and the true baselines seem like odd things to assert about the context. These baselines provide evidence that the task should be interpreted as a truth-value judgment task. We therefore interpret participants' behavior with *must p* in experiments 1 and 2 as follows: in the one-sentence task (experiment 1) more people agreed with *must p* because more people interpreted the question as 'Would you conclude that *p* in this context?', whereas in the multiple-sentence task (experiment 2) fewer participants agreed with *must p* because they were induced to interpret the task as a truth-value task.

We take finding C—people are led to conclude that an event happened independently from their estimation of the objective probability of the event—as consistent with findings in the psychology of decision making which suggest that people's decisions to believe in an event are determined based on some subjective heuristics and biases more than on the objective probability of the event (e.g. Tversky & Kahneman 1974, 1992, Gigerenzer et al. 2011, Preuschoff et al. 2013, Gonzalez 2016).

We take finding D—people are more induced to conclude that an event happened if this event is the most plausible explanation for another event than if the event is just supported by statistical reasoning—as consistent with findings in the psychology of reasoning that people tend to overestimate the likelihood of a good and simple explanation (Chapman & Chapman 1967, 1969, Ross et al. 1975, Anderson et al. 1980, Koehler 1991, Brem & Rips 2000, Lombrozo 2007, 2012). Indeed, our participants were much more inclined to endorse the statement *Bill must not have won the raffle* (but also *It is certain that Bill did not win the raffle*) when the event of Bill having not won the lottery was a good explanation for the fact that he looked disappointed the day after the lottery drawing than when the occurrence of the same event was supported only by statistical reasoning based on the number of chances of winning.

We also take our findings here to be in line with previous findings about epistemic *must* reported in the literature. First, Degen et al. (2019) designed a battery of experiments meant to test the meaning of several English and German expressions—including English *must* and German *muss*—with respect to the strength of speaker commitment. Overall, they found that by uttering *must p* and *muss p* speakers express a commitment to the truth of *p* that is weaker than that expressed by uttering just *p* but stronger than that expressed by uttering *probably p*. These findings can be accounted for by the evidential hypothesis by assuming (as we do here) that presenting *p* as a conclusion expresses a weaker commitment to the truth of *p* than presenting *p* as an established fact, but a stronger commitment than presenting *p* as just probable.

Second, Del Pinal and Waldon (2019) presented two sets of experiments meant to assess the logical hypothesis, the probabilistic hypothesis, and the weak logical hypothesis. In their first set of experiments, they found that participants rated conjoined statements like *must p but I don't know for sure that p* more acceptable than *it is certain that p but I don't know for sure that p*, which suggests that *must p* followed by an explicit denial of knowledge/certainty in *p* is felt to be less contradictory than *it is certain p* followed by an explicit denial of knowledge in *p*, contrary to what the hyperbolic logical hypothesis predicts. In their second set of experiments, they found that in situations where a speaker says either *must p* or *it is almost certain that p* and it turns out that *p* is false, it is harder for those speakers to deny that they were wrong in saying *must p* than in saying *it is almost certain that p*, which suggests that *must p* is a stronger statement than *almost certain that p* (= 'it is highly probable that *p*'), contrary to what the probabilistic hypothesis predicts. Del Pinal and Waldon take these findings as supporting the weak logical *must* hypothesis—epistemic *must* expresses certainty relative to some factual evidence and some normality assumption—which in their interpretation does not seem to differ empirically from the evidential hypothesis (see n. 7 for a discussion of the interpretive problems raised by this hypothesis). The findings from these experiments can be accounted for by the evidential hypothesis: the first set of findings is accounted for by assuming that people do not need to be sure about an event to conclude that it happened, and the second set is accounted for by assuming that people have a hard time denying that they were wrong in concluding that an event happened once they discover that such an event did not happen.

Even though we have argued that the evidential hypothesis offers a better account of our findings than the (hyperbolic) logical hypothesis and the probabilistic hypothesis do, nonetheless the present results do not settle the debate about the meaning of epistemic *must p*. In fact, we suggest that the hyperbolic logical view of *must p* advanced by von Fintel and Gillies (2021) is not completely wrong. Here is our speculation.

The logical view captures the original meaning of *must p* well, where the word *must* conveys the meaning of a necessity operator such that epistemic *must p* means roughly '*p* is a necessary conclusion'. In addition, strong expressions tend to be overused for rhetorical purposes, which, over time, can trigger an INFLATIONARY EFFECT leading to a devaluation of the expression (cf. Keller 1989, Haspelmath 1999, Dahl 2001, Deo 2015). For example, emphatic negation constructions tend to develop into nonemphatic constructions ('Jespersen's cycle'): constructions like the French *ne ... pas* initially express emphatic negation and become over time the standard nonemphatic means of expressing negation in the language. We speculate that a similar inflationary effect has occurred in the case of epistemic *must p*. For rhetorical purposes, English speakers tend to use the *must p* statement beyond the restricted boundaries of logical inferences to include nonlogical inferences they feel strongly confident about. Over time, with the increase of such rhetorical/emphatic uses, at least in everyday communication, *must* loses its status as a marker of the special case of logical inferences and becomes a generic marker of inference used by speakers to mark conclusions whose truth they are strongly convinced of.

Nonetheless, this process of reanalysis leading to the use of *must* and *deve* as evidential markers is not complete, because the necessity component of the meaning of *must* (its logical value) may not have been completely eliminated for every speaker; indeed, the original meaning of the word can be retrieved at any time when someone utters an epistemic *must p* statement. For example, one can offer the following reply to any of the participants in our experiment 3 who endorsed *Bill must not have won the raffle* in the explanatory context.

(17) Forcing *must* to be interpreted as a logical statement:
It is not true that Bill must not have won the raffle because he looks disappointed.

In this reply, the speaker is focusing on the word *must* and in doing so reveals what the word literally means: 'it is necessarily the case that'. A similar move can be made in Italian.

We suggest that the evidential sense of *must* and *deve* is encoded in the current semantics of the words as one of their established senses. We hypothesize that *must* and *deve* are currently polysemous between two established epistemic senses (in addition to the deontic sense and the other nonepistemic senses encoded in these words): the original epistemic necessity sense, and the inferential sense derived from it by extension through a process of rhetorical devaluation. Thus, we believe the hyperbolic logical hypothesis is right in identifying a process of exaggeration as the key component triggering the semantic extension that created the inferential sense from the epistemic necessity sense, but that this hypothesis is wrong in assuming that speakers typically use epistemic *must* and *deve* with their original meaning to communicate in an exaggerated fashion that an event is certain. Furthermore, we propose to reinterpret Lyons's (1977) distinction between an objective and a subjective use of epistemic *must* (see also Papafragou 2006, Yatsushiro et al. 2022) in our terms as a distinction between the use of *must* as a necessity operator and the use of *must* as an inferential marker.

From a typological perspective, the question arises naturally as to whether the semantic extension from the epistemic necessity meaning to the inferential meaning is specific to the English *must* and the Italian *deve* or is a general property of any expression encoding the concept of epistemic necessity across languages. Plausibly, every such expression is bound to undergo a process of devaluation if its frequency of use increases: rarely in daily communication are we in the position of presenting a conclusion of our reasoning as certain. Therefore, an increased use of an epistemic necessity operator for rhetorical purposes would have as a natural effect a weakening of the original meaning. Indeed, since 'natural language has no practical use for an epistemic necessity operator' (Westmoreland 1998:54), expressions that originally encode such operators can be profitably used only if they are weakened to be generic markers of inference: any speech act originally communicating something like 'some evidence available to me makes it necessary that this event occurred' is bound to be reinterpreted roughly as 'some evidence available to me compels me to conclude (= think) that this event occurred'.

4. CONCLUSION. In this work we have investigated the meaning of *must* and *deve* in their epistemic use. We have offered evidence from four experiments suggesting that people typically interpret these words as expressing the meaning of an inferential marker. Moreover, we have shown that whether speakers would present themselves as having concluded that an event happened depends not on the objective probability of that event—whether the event is certain or highly probable—but on its subjective probability estimated on the basis of mental biases about the nature of that event. Interestingly, these findings confirm the more general picture about how people form beliefs, as evidenced from findings in the psychology of reasoning: people tend to violate laws of logic or probability in determining their confidence in the possible occurrence of an event and rely, instead, on heuristics and cognitive biases defining the subjective probability of that event.

We have also speculated that this meaning of *must* and *deve* is derived from their overuse as markers of epistemic necessity: the original meaning indicating logical conclusions (i.e. certainties) is weakened/bleached to indicate just conclusions (i.e. opinions) after reiterated and implausible exaggerated uses of the words to refer to very likely events as certain. We further suggested that, crosslinguistically, any expression encoding an epistemic necessity operator is destined to undergo such a process of semantic extension if its use increases: reiterated attempts by speakers to present an event as certain based on their reasoning would naturally lead listeners to reinterpret the marker of epistemic necessity with the weaker and more plausible communicative import of a marker of inference.

REFERENCES

Anderson, Craig A.; Mark R. Lepper; and Lee Ross. 1980. Perseverance of social theories: The role of explanation in the persistence of discredited information. *Journal of Personality and Social Psychology* 39.1037–49. DOI: 10.1037/h0077720.

Brem, Sarah K., and Lance J. Rips. 2000. Explanation and evidence in informal argument. *Cognitive Science* 24.573–604. DOI: 10.1016/S0364-0213(00)00033-1.

Carnap, Rudolf. 1964. *Meaning and necessity: A study in semantics and modal logic*. Chicago: University of Chicago Press.

Chapman, Loren J., and Jean P. Chapman. 1967. Genesis of popular but erroneous psychodiagnostic observations. *Journal of Abnormal Psychology* 72.193–204. DOI: 10.1037/h0024670.

Chapman, Loren J., and Jean P. Chapman. 1969. Illusory correlation as an obstacle to the valid use of psychodiagnostic signs. *Journal of Abnormal Psychology* 74.271–80. DOI: 10.1037/h0027592.

Dahl, Östen. 2001. Inflationary effects in language and elsewhere. *Frequency and the emergence of linguistic structure*, ed. by Joan Bybee and Paul Hopper, 471–80. Amsterdam: John Benjamins.

Davidson, Kathryn. 2020. Is 'experimental' a gradable predicate? *North East Linguistic Society (NELS)* 50.125–44.

Degen, Judith; Andreas Trotzke; Gregory Scontras; Eva Wittenberg; and Noah D. Goodman. 2019. Definitely, maybe: A new experimental paradigm for investigating the pragmatics of evidential devices across languages. *Journal of Pragmatics* 140. 33–48. DOI: 10.1016/j.pragma.2018.11.015.

Del Pinal, Guillermo. 2021. Probabilistic semantics for epistemic modals: Normality assumptions, conditional epistemic spaces and the strength of *must* and *might*. *Linguistics and Philosophy* 45.985–1026. DOI: 10.1007/s10988-021-09339-6.

Del Pinal, Guillermo, and Brandon Waldon. 2019. Modals under epistemic tension. *Natural Language Semantics* 27.135–88. DOI: 10.1007/s11050-019-09151-w.

Deo, Ashwini. 2015. Diachronic semantics. *Annual Review of Linguistics* 1.179–97. DOI: 10.1146/annurev-linguist-030514-125100.

Douven, Igor. 2017. Abduction. *The Stanford encyclopedia of philosophy* (Summer 2017 edn.), ed. by Edward N. Zalta. Online: https://plato.stanford.edu/archives/sum2017/entries/abduction/.

Drubig, Hans Bernhard. 2001. On the syntactic form of epistemic modality. Tübingen: University of Tübingen, ms.

Eckardt, Regine, and Andrea Beltrama. 2019. Evidentials and questions. *Empirical Issues in Syntax and Semantics* 12.121–55. Online: http://www.cssp.cnrs.fr/eiss12/eiss12_eckardt-and-beltrama.pdf.

Faller, Martina T. 2002. *Semantics and pragmatics of evidentials in Cuzco Quechua*. Stanford, CA: Stanford University dissertation.

Faller, Martina T. 2019. The discourse commitments of illocutionary reportatives. *Semantics and Pragmatics* 12:8. DOI: 10.3765/sp.12.8.

Foppolo, Francesca; Maria Teresa Guasti; and Gennaro Chierchia. 2012. Scalar implicatures in child language: Give children a chance. *Language Learning and Development* 8.365–94. DOI: 10.1080/15475441.2011.626386.

Frana, Ilaria, and Paula Menéndez-Benito. 2019. Evidence and bias: The case of the evidential future in Italian. *Proceedings of Semantics and Linguistic Theory (SALT)* 29.727–47. DOI: 10.3765/salt.v29i0.4629.

Giannakidou, Anastasia, and Alda Mari. 2016. Epistemic future and epistemic MUST: Nonveridicality, evidence, and partial knowledge. *Mood, aspect, modality revisited: New answers to old questions*, ed. by Joanna Błaszczak, Anastasia Giannakidou, Dorota Klimek-Jankowska, and Krzysztof Migdalski, 75–117. Chicago: University of Chicago Press. DOI: 10.7208/chicago/9780226363660.003.0003.

Giannakidou, Anastasia, and Alda Mari. 2018. A unified analysis of the future as epistemic modality. *Natural Language & Linguistic Theory* 36.85–129. DOI: 10.1007/s11049-017-9366-z.

Gibson, Edward, and Evelina Fedorenko. 2013. The need for quantitative methods in syntax and semantics research. *Language and Cognitive Processes* 28.88–124. DOI: 10.1080/01690965.2010.515080.

Gigerenzer, Gerd; Ralph Hertwig; and Thorsten Pachur. 2011. *Heuristics: The foundation of adaptive behavior*. New York: Oxford University Press. DOI: 10.1093/acprof:oso/9780199744282.001.0001.

Gonzalez, Cleotilde. 2016. Decision-making: A cognitive science perspective. *The Oxford handbook of cognitive science*, ed. by Susan E. F. Chipman, 249–64. Oxford: Oxford University Press. DOI: 10.1093/oxfordhb/9780199842193.013.6.

Goodhue, Daniel. 2017. *Must* ϕ is felicitous only if ϕ is not known. *Semantics and Pragmatics* 10:14. DOI: 10.3765/sp.10.14.

Harman, Gilbert H. 1965. The inference to the best explanation. *The Philosophical Review* 74.88–95. DOI: 10.2307/2183532.

Haspelmath, Martin. 1999. Why is grammaticalization irreversible? *Linguistics* 37.1043–68. DOI: 10.1515/ling.37.6.1043.

Hintikka, Kaarlo Jaakko Juhani. 1962. *Knowledge and belief: An introduction to the logic of the two notions*. Ithaca, NY: Cornell University Press.

Izvorski, Roumyana. 1997. The present perfect as an epistemic modal. *Proceedings of Semantics and Linguistic Theory (SALT)* 7.222–39. DOI: 10.3765/salt.v7i0.2795.

Jackendoff, Ray. 1972. *Semantic interpretation in generative grammar*. Cambridge, MA: MIT Press.

Jasbi, Masoud; Brandon Waldon; and Judith Degen. 2019. Linking hypothesis and number of response options modulate inferred scalar implicature rate. *Frontiers in Psychology* 10:189. DOI: 10.3389/fpsyg.2019.00189.

Karttunen, Lauri. 1972. Possible and must. *Syntax and semantics, vol. 1*, ed. by John Kimball, 1–20. New York: Brill. DOI: 10.1163/9789004372986_002.

Katsos, Napoleon, and Dorothy V. M. Bishop. 2011. Pragmatic tolerance: Implications for the acquisition of informativeness and implicature. *Cognition* 120.67–81. DOI: 10.1016/j.cognition.2011.02.015.

Keller, Rudi. 1989. Invisible-hand theory and language evolution. *Lingua* 77.113–27. DOI: 10.1016/0024-3841(89)90011-9.

Koehler, Derek J. 1991. Explanation, imagination, and confidence in judgment. *Psychological Bulletin* 110.499–519. DOI: 10.1037/0033-2909.110.3.499.

Kratzer, Angelika. 1981. The notional category of modality. *Words, worlds, and contexts: New approaches in word semantics*, ed. by Hans-Jürgen Eikmeyer and Hannes Rieser, 38–74. Berlin: De Gruyter. DOI: 10.1515/9783110842524-004.

Kratzer, Angelika. 1991. Modality. *Semantics: An international handbook of contemporary research*, ed. by Arnim von Stechow and Dieter Wunderlich, 639–50. Berlin: De Gruyter. DOI: 10.1515/9783110126969.7.639.

Kratzer, Angelika. 2012. *Modals and conditionals: New and revised perspectives*. Oxford: Oxford University Press. DOI: 10.1093/acprof:oso/9780199234684.001.0001.

Lassiter, Daniel. 2016. *Must*, knowledge, and (in)directness. *Natural Language Semantics* 24.117–63. DOI: 10.1007/s11050-016-9121-8.

Lassiter, Daniel. 2017. *Graded modality: Qualitative and quantitative perspectives*. Oxford: Oxford University Press. DOI: 10.1093/oso/9780198701347.001.0001.

Lewis, David. 1968. *On the plurality of worlds*. Oxford: Basil Blackwell.
Lombrozo, Tania. 2007. Simplicity and probability in causal explanation. *Cognitive Psychology* 55.232–57. DOI: 10.1016/j.cogpsych.2006.09.006.
Lombrozo, Tania. 2012. Explanation and abductive inference. *The Oxford handbook of thinking and reasoning*, ed. by Keith J. Holyoak and Robert G. Morrison, 260–76. Oxford: Oxford University Press. DOI: 10.1093/oxfordhb/9780199734689.013.0014.
Lyons, John. 1977. *Semantics*, vol. 2. Cambridge: Cambridge University Press.
Mandelkern, Matthew. 2019. What 'must' adds. *Linguistics and Philosophy* 42.225–66. DOI: 10.1007/s10988-018-9246-y.
Marty, Paul; Emmanuel Chemla; and Jon Sprouse. 2020. The effect of three basic task features on the sensitivity of acceptability judgment tasks. *Glossa: a journal of general linguistics* 5(1):72. DOI: 10.5334/gjgl.980.
Matthewson, Lisa; Henry Davis; and Hotze Rullmann. 2007. Evidentials as epistemic modals: Evidence from St'át'imcets. *Linguistic Variation Yearbook* 7.201–54. DOI: 10.1075/livy.7.07mat.
McCready, Elin. 2010. Varieties of conventional implicature. *Semantics and Pragmatics* 3:8. DOI: 10.3765/sp.3.8.
Murray, Sarah E. 2017. *The semantics of evidentials*. Oxford: Oxford University Press. DOI: 10.1093/oso/9780199681570.001.0001.
Nuyts, Jan. 2016. Analyses of the modal meanings. *The Oxford handbook of modality and mood*, ed. by Jan Nuyts and Johan van der Auwera, 31–49. Oxford: Oxford University Press. DOI: 10.1093/oxfordhb/9780199591435.013.1.
Palmer, Frank Robert. 2001. *Mood and modality*. Cambridge: Cambridge University Press.
Papafragou, Anna. 2006. Epistemic modality and truth conditions. *Lingua* 116.1688–1702. DOI: 10.1016/j.lingua.2005.05.009.
Pietrandrea, Paola. 2005. *Epistemic modality: Functional properties and the Italian system*. Amsterdam: John Benjamins. DOI: 10.1075/slcs.74.
Preuschoff, Kerstin; Peter N. C. Mohr; and Ming Hsu. 2013. Decision making under uncertainty. *Frontiers in Neuroscience* 7:218. DOI: 10.3389/fnins.2013.00218.
Ross, Lee D.; Mark R. Lepper; and Michael Hubbard. 1975. Perseverance in self-perception and social perception: Biased attributional processes in the debriefing paradigm. *Journal of Personality and Social Psychology* 32(5).880–92. DOI: 10.1037/0022-3514.32.5.880.
Ross, Lee D.; Mark R. Lepper; Fritz Strack; and Julia Steinmetz. 1977. Social explanation and social expectation: Effects of real and hypothetical explanations on subjective likelihood. *Journal of Personality and Social Psychology* 35.817–29. DOI: 10.1037/0022-3514.35.11.817.
Skordos, Dimitrios, and Anna Papafragou. 2016. Children's derivation of scalar implicatures: Alternatives and relevance. *Cognition* 153.6–18. DOI: 10.1016/j.cognition.2016.04.006.
Sprouse, Jon, and Diogo Almeida. 2017. Design sensitivity and statistical power in acceptability judgment experiments. *Glossa: a journal of general linguistics* 2(1):14. DOI: 10.5334/gjgl.236.
Squartini, Mario. 2001. The internal structure of evidentiality in Romance. *Studies in Language* 25.297–334. DOI: 10.1075/sl.25.2.05squ.
Stephenson, Tamina. 2007. Judge dependence, epistemic modals, and predicates of personal taste. *Linguistics and Philosophy* 30.487–525. DOI: 10.1007/s10988-008-9023-4.
Stone, Matthew. 1994. The reference argument of epistemic *must*. *IWCS-1 '94: Proceedings of the First International Conference on Computational Semantics*, 181–90.
Swanson, Eric. 2011. How not to theorize about the language of subjective uncertainty. *Epistemic modality*, ed. by Egan Andy and Brian Weatherson, 249–69. Oxford: Oxford University Press. DOI: 10.1093/acprof:oso/9780199591596.003.0009.
Syrett, Kristen. 2015. Experimental support for inverse scope readings of finite-clause-embedded antecedent-contained deletion sentences. *Linguistic Inquiry* 46.579–92. DOI: 10.1162/LING_a_00194.

SYRETT, KRISTEN, and ADRIAN BRASOVEANU. 2019. An experimental investigation of the scope of object comparative quantifier phrases. *Journal of Semantics* 36.285–315. DOI: 10.1093/jos/ffy019.

SYRETT, KRISTEN, and TODOR KOEV. 2015. Experimental evidence for the truth conditional contribution and shifting information status of appositives. *Journal of Semantics* 32.525–77. DOI: 10.1093/jos/ffu007.

TONHAUSER, JUDITH; DAVID BEAVER; CRAIGE ROBERTS; and MANDY SIMONS. 2013. Toward a taxonomy of projective content. *Language* 89.66–109. DOI: 10.1353/lan.2013.0001.

TVERSKY, AMOS, and DANIEL KAHNEMAN. 1974. Judgment under uncertainty: Heuristics and biases: Biases in judgments reveal some heuristics of thinking under uncertainty. *Science* 185.1124–31. DOI: 10.1126/science.185.4157.1124.

TVERSKY, AMOS, and DANIEL KAHNEMAN. 1992. Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty* 5.297–323. DOI: 10.1007/BF00122574.

VELTMAN, FRANK. 1996. Defaults in update semantics. *Journal of Philosophical Logic* 25.221–61. DOI: 10.1007/BF00248150.

VON FINTEL, KAI, and ANTHONY S. GILLIES. 2010. *Must* … stay … strong! *Natural Language Semantics* 18.351–83. DOI: 10.1007/s11050-010-9058-2.

VON FINTEL, KAI, and ANTHONY S. GILLIES. 2021. Still going strong. *Natural Language Semantics* 29.91–113. DOI: 10.1007/s11050-020-09171-x.

WALDON, BRANDON, and JUDITH DEGEN. 2020. Modeling behavior in truth value judgment task experiments. *Proceedings of the Society for Computation in Linguistics 2020*, 238–47. Online: https://aclanthology.org/2020.scil-1.29.

WESTMORELAND, ROBERT RALPH. 1998. *Information and intonation in natural language modality*. Bloomington: Indiana University dissertation.

YATSUSHIRO, KAZUKO; TUE TRINH; MARZENA ŻYGIS; STEPHANIE SOLT; ANTON BENZ; and MANFRED KRIFKA. 2022. Certainly but not certain: The expression of subjective and objective probability. *Glossa: a journal of general linguistics* 7(1). DOI: 10.16995/glossa.5847.

ZEHR, JÉRÉMY, and FLORIAN SCHWARZ. 2018. Returning to non-entailed presuppositions again. *ZAS Papers in Linguistics* 61.463–80. DOI: 10.21248/zaspil.61.2018.507.

[gricciardi@g.harvard.edu]
[rryskin@ucmerced.edu]
[egibson@mit.edu]

[Received 22 July 2021;
revision invited 12 February 2022;
revision received 30 September 2022;
accepted pending revisions 28 February 2023;
revision received 3 April 2023;
accepted 10 April 2023]