Epistemic 'must p' is literally a strong statement

Intuitively, in a sentence like 'Bill must not have won the raffle,' 'must p' conveys that the speaker is inferring 'p' ('Bill has not won the raffle'). However, the strength of this inference has been debated at least since the 1970s in the theoretical literature (see a.o. Kartunnen, 1972; Lyons, 1977; Kratzer, 1981; von Fintel & Gillies [F&G], 2010, 2018) and more recently in the experimental literature (Lassiter, 2016; Del Pinal & Waldon, 2019; Degen et al., 2019). Here, we focus on Lassiter's (2016; L16) assessment of F&G (2010)'s hypothesis that 'must p' defines 'p' as a *deductive/necessary conclusion* and L16's hypothesis that 'must p' defines 'p' as an *inductive/probabilistic conclusion*. In this work, we report a replication of L16 and 2 follow-up studies. Our results support F&G's hypothesis regarding the semantics of 'must'.

Replication of L16 (Expt 1). Participants on Amazon's MTurk (N = 180) read a context passage (1) in which Bill has a 1/1000 chance of winning a raffle. They then responded whether they agreed or disagreed with a single statement out of 9 possible statements (between-subjects) including: 'Bill must not have won the raffle', 'It is certain that Bill did not win the raffle', 'We know that Bill did not win the raffle'. L16 found (Fig.1-left) that the agreement rate for the 'must' statement (58%) was significantly higher than for 'certain' (25%) and 'know' (21%). We replicate this qualitative pattern (Fig.1-middle): the proportion for 'must' (28%) is significantly higher than for 'certain' (9%) and for 'know' (8%; $p<.01$), though agreement rates were lower overall. Assuming that 'certain' and 'know' express the speaker's full confidence in the truth of 'p', L16 interpreted this finding as supporting the hypothesis that 'must p' expresses less than full confidence in 'p'. However, it is unclear whether participants interpreted the "agree" /"disagree" task with respect to the truth-conditions (i.e., literal meaning) of the statement. They also endorsed statements like 'Bill did not win the lottery', which is underdetermined by the context. We therefore hypothesize that some participants may guess that the task is judging *the likelihood of the described events in the context* - perhaps a simpler language task - rather than *the truth value of the statement.* Thus, we predicted that, in the presence of unambiguously true or false statements, participants would be more likely to converge on the intended truth-conditional interpretation of the task. In **Expt 2**, each participant (N = 180) read the same lottery context but rated all three experimental items plus four clearly true and two clearly false sentences (see (2)). The agreement rate for 'must' was decreased relative to E1 (6%) and did not differ from 'certain' (4%) or 'know' (3%) (Fig. 1-right), whereas 'probable' was endorsed by 90% (see Fig. 2). In **Expt 3,** we provided our participants (N = 2000, planned via power analysis) with either 1) one statement (as in E1; **Baseline** in Fig 3); 2) examples of statements to which the answer should be 'agree' or 'disagree' (**Examples** in Fig 3); 3) an unambiguously false statement to evaluate before 'must' (**One false** in Fig 3); or 4) an unambiguously true additional statement to evaluate before 'must' (**One true** in Fig 3). As predicted, participants' agreement rate for 'must' was lower for the **One true** (8%, $p<0.05$) and (numerically) for the **One false** (10%, $p=0.12$) conditions relative to **Baseline** (14%) (see Fig.3).

We showed that modifications of L16 which encourage participants to understand the task as a truth-value judgment task lead to lower endorsement rates for statements with epistemic 'must': people were probably interpreting L16's task as a plausibility judgment task, instead of the intended truth-value judgment task. When task effects are removed, the semantics of 'must p' are more closely aligned with 'certain' or 'know', in contrast to L16, but consistent with F&G's account. However, in everyday usage, 'must' is often used to convey a "weaker" meaning. We speculate that speakers often use 'must' hyperbolically to persuade listeners that their conclusion follows deductively from the evidence ('She responded so quickly; she must have been waiting for my email.'). In light of this, for listeners in everyday communication, the pragmatically enriched meaning of 'must' may be closer to 'probably'.

(1) Experimental 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

(2) Experiment 2 conditions and their labels in the graphs

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) *There is a slight chance that Bill won the raffle* ("chance");
(e) *It is highly probable that Bill did not win the raffle* ("probable");
(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")

(3) Experiment 3: the four conditions

1. 'Baseline': as in Lassiter; 2. 'Examples': task explained through explanatory paragraph; 3. 'One False': preceding obviously false statement; 4. 'One True': preceding obviously true statement

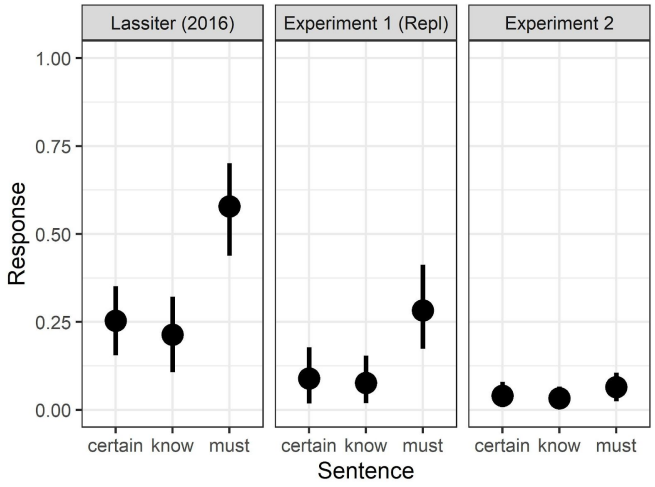

Figure 1. Results of Lassiter (2016), our Experiment 1 (Replication) and Experiment 2 for the three test items.

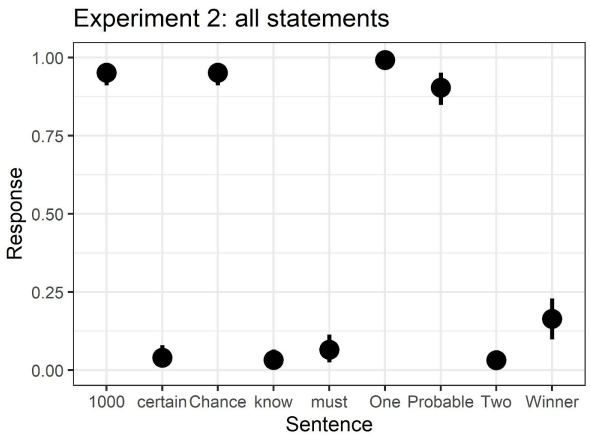

Figure 2. Results of Experiment 2 for all items

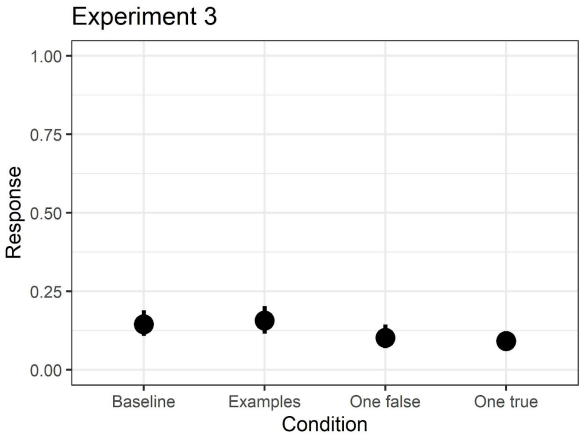

Figure 3. Results of Experiment 3