**Responses to reviewers’ comments**:

**Reviewer 1**

[1] i. l.801ff.: On the marginal possibility of 'must/müssen' in direct evidence contexts:

Looking at the bars in Fig3, it seems that the proposition especially of English 'must' is really really small with direct perception, so couldn't this be explained away as mere performance errors (tiredness, lack of attention, confusion). Perhaps you would want to stress again (in a footnote) that each individual participant did not see so many items to induce tiredness?

Also, intuitively, the chance of performance errors seems smaller in forced production than in more passive comprehension. Again, it would be instructive to hear a sentence or two on why the authors find the marginal occurrence of 'must' with direct evidence worth mentioning.

***Response: We thank the reviewer for this suggestion. We have now included such a footnote (Footnote 5 on p. 13).***

[2] It would also be helpful to see a comment on why 'presupposition accommodation' does not seem to help in accounting for the occurrence of 'must' with direct evidence contexts. Looking at the stories, I couldn't really think of a PLAUSIBLE story to license presupposition accommodation, such that the relevant statement is based on indirect evidence after all. The only thing that comes to mind is a Cartesian distrust in the accuracy of one's own senses, and in particular one's vision.

But I am not sure that this is what people actually did. The authors could point out (in a footnote) that presupposition accommodation is not as easy a way out as in other cases of apparent presupposition violations.

***Response: We have addressed this in the same footnote.***

[5] ii. this point regards the labelling of on context in the appendix:

Ai.2 (rain - sound on the roof): I am not sure that this is accurately labelled as indirect evidence, as it is directly related to an auditory sense stimulus. It is only indirect in the sense that one cannot see the rain, but that should not matter for auditory perception (see the system of evidence types in (6), in which 'auditory perception' is labelled as 'direct'. The same might be said for haptic evidence in Aii.2. I guess at this point, the authors cannot see this experiment any longer, but I would find it reassuring to know that the statistics do not change if Ai.2 and Aii.2 are labelled as 'direct evidence' instead of indirect evidence.

At the very least, the choice of the label 'indirect' should be motivated for these two cases.

*Response: We spent some discussing those particular items amongst ourselves and ended up categorizing them as indirect for the following reason: what should be relevant for the evidence to be categorized as direct is whether or not the evidence is direct evidence* for p *(that it is raining), and not simply whether the evidence is perceptual. If the only bar for directness is whether evidence is perceptual, then hallucinating the sound of rain should be ‘direct’ evidence for it raining. Similarly, the haptic evidence in A.2.2 might be misleading – maybe without my knowing it, the cold cup is only an externally cold thermos cup containing hot coffee.*

*This is all to say that we explicitly stacked the deck against ourselves – if we instead label these two pieces of evidence ‘direct’, then we also see more ‘must’ uses with direct evidence in production. We have included the graph that demonstrates this below. In the production analysis, all the results remain qualitatively the same with the following exceptions: the effect of evidence strength on English ‘must’ changes from significant to marginal (without changing sign) and the effect of evidence directness for German ‘vermutlich’ changes from non-significant to significant (without changing sign).*



*Similarly, in comprehension, if we label the two pieces of evidence in question as direct, we observe more directness inferences in listeners who observe ‘must/muss p’. Again we have included the graph that demonstrates this below. The qualitative results from the statistical analysis remain the same.*

**

*We have addressed this issue in Footnote 9.*

**Reviewer 2**

[6] The paper makes an experimental contribution to the debate over epistemic modals. I found the experimental paradigm novel, well executed, and useful for future research. Though, I do agree with an earlier referee's assessment that the results really show only "differences or similarities in the use conditions and some interpretive effects associated with [lexical expressions that convey different strengths of speaker commitment], which may or may not be correlated with underlying semantic differences." In fact, the greatest shortcoming of the paper is the lack of a sustained consideration of pragmatic factors that might figure into the interpretation of these results. For example, consider the claim at the end of the paper, that "the results also suggested that speaker commitment is lower for *must* than for the bare form. This is at odds with von Fintel & Gillies’ claim that '[s]peakers who say *must* *F* are just as strongly committed to the prejacent as those who assert *F* by itself' (von Fintel & Gillies 2010: 30)." A serious discussion of pragmatic factors would weaken the claim that the finding concerning *must* is at odds with vF & G’s claim, for on most pragmatic frameworks (e.g., (neo-)Gricean, Relevance Theoretic, etc.) there is consideration given to the processing costs associated with an expression. So, for example, on the relevance theoretic picture, the interpretation procedure aims for an interpretation of an utterance that is optimally relevant, where an interpretation is more relevant the more positive cognitive effects it engenders, and less relevant the more processing costs required to arrive at it. Therefore, even if *must F* conveys as strong a commitment as the bare form of *F*, the addition of *must* increases processing cost, thereby causing interpreters to look for some additional information conveyed by the utterance to offset these costs and arrive at an optimally relevant interpretation. That additional information could be a hedge on the basic commitment that would have been conveyed by the bare form. Similarly, a Gricean picture would emphasize that a speaker could have said something more perspicuous by using the bare form, and therefore must imply something more through the addition of the superfluous *must*. Neither of these accounts entails that the addition of *must* serves toconvey anything weaker than the bare form without the addition of *must* would have at the level of expression meaning. (I have focused on the interpretation/hearer’s side, but, of course, these and other pragmatic theories see a tight connection between how speakers interpret utterances and how speakers use utterances to communicate, so these accounts would also be useful in thinking about the results regarding speaker commitments.) We see the same basic pragmatic phenomena at work in the case of so-called transparent belief reports, such as “I believe the keys are in the car”—or, more to the point, perhaps, “I just know the keys are in the car”—where the commitment seems weaker than the bare claim that “the keys are in the car”. (See, for example, Kauppinen, 2010, “The pragmatics of transparent belief reports,” or Ifantidou, 1994, *Evidentials and Relevance*.)

*Response: XXX*