Responses to Reviewer's Comments:

[1] "[…] on a slightly different presentational mode of the empirical data, the paper might at least serve to evaluate the prominent proposal by von Fintel & Gillies (2010), according to which modal epistemic *must* is restricted to indirect evidential contexts. The question of what does and what doesn't count as an indirect evidential context is left open by von Fintel & Gillies (2010), and this is where the current paper could have made a valuable contribution, as the authors test each modal/evidential expression in various evidential contexts (direct perception vision, perception olfactory, hearsay, indirect inferencing, ...)."
Response: ***Thank you very much for pointing this out. We have substantially modified our paper to report on detailed analyses of evidence strength and type. Our results suggest that von Fintel & Gillies are mostly correct in claiming that*** **must** ***resists direct evidential contexts, but our results also add some nuance to this claim by demonstrating that the restriction is not absolute:*** **must** ***sometimes allows direct evidential contexts.***

[2] "[…] the authors might want to consider (re)phrasing some of their findings with reference to a ***qualitative characterization*** of different evidential contexts. At the very least, they should motivate their choice for a purely quantitative measure of evidence strength. And they should be more articulate in stressing that the reported data do not have a direct bearing on existing theoretical approaches."
Response: ***We now report both quantitative (i.e., evidence strength) and qualitative (i.e., evidence type) characterizations of evidential contexts, thereby more directly connecting our results with the relevant claims from the literature.***

[3] "[…] the authors formulate a rather modest, but correct assessment of their contribution to theory formation: 'The paradigm provides a starting point for approaching theoretical debates on the nature of evidential expressions empirically'. I would insert 'potential' before 'starting point'."
Response: ***Done.***

[4] "What are the blind spots in theoretical debates? How is the underlying semantic or pragmatic status of such evidential devices tied to commitment strength? Why would that be important from a theoretical perspective?"
Response: ***We added the following paragraph to the relevant passage where we talk about blind spots. We would like to postpone the detailed discussion of these 'blind spots' to the individual sections where we actually deal with the debates we are alluding to here.***
"[…] and we will argue that our experiments on different degrees of speaker commitment in the domain of these evidential devices address obvious blind spots in the theoretical debates. More specifically, both empirical phenomena (epistemic *must* and epistemic discourse particles) are usually only discussed with regard to their theoretical status as semantic and/or pragmatic elements, and the literature does not provide a detailed picture of the evidential contexts these elements can actually be used in. For *must*, our data shed some light on what kind of evidential contexts can count as the relevant indirect contexts where the epistemic use of *must* is felicitous. For discourse particles, we explore whether particles like *wohl* differ in their compatibility with different evidential circumstances, compared to closely related elements like synonymous higher (i.e., speaker-oriented) adverbs."

[5] "[…] The rhetoric is even stronger on p.4: 'The experimental data we present below cannot decide between ***all*** of these accounts'; and later on p.15: 'Although our experimental data cannot

decide ***fully*** between the various theoretical accounts on [...] epistemic *must* [...] and discourse particles'. These phrasing give rise to implicatures that are too strong, and they stand in stark contrast to what the paper actually achieves in theoretical terms. They suggest that the experimental data shed at least some light on the (in)validity of existing accounts, but I fail to see how. *The authors should adjust these passages in line with the weaker statement at the outset.*"

Response: ***We revised these passages as follows:***

"The experimental data we present below cannot decide between these theoretical accounts. […] Although our experimental data cannot decide between the various theoretical accounts on […]"

[6] "[…] This problem or misunderstanding resurfaces in the conclusion (p.16), where the authors claim that 'our new experimental paradigm thus illustrates a new approach to detect and define meaning differences [sic!] in the lexical inventory of evidentials'. *For the time being, I would be more careful and talk about differences or similarities in the use conditions and some interpretive effects associated with such items, which may or may not be correlated with underlying semantic differences.*"

Response: ***We agree with the reviewer and rephrased as follows:***

"Our new experimental paradigm thus illustrates a new approach that focuses on differences and similarities in the use conditions of the lexical inventory of evidentials across languages. In doing so, we provide a starting point for adding a use-oriented view to theoretical debates on the nature of evidential expressions and highlight the importance of an experimentally-driven perspective in this context."

[7] "I wonder why the authors did not opt for alternative experimental paradigms (eg. acceptability ratings in context), which would shed more light on the theoretical analyses from the literature, say by testing for presupposition violation or implicature cancellation? *The authors should comment on why the choice of this new method over more conventional approaches would be advantageous, over and beyond delivering novel experimental data – the theoretical significance of which is unclear*. I find this type of discussion increasingly important, given the current lay of the land in experimental semantics and pragmatics."

Response: ***We added the following clarification to our introduction.***

"In particular, we introduce a methodology for exploring the impact of different evidential circumstances on the use of modal evidentials, epistemic discourse particles, and statements with no evidential markers at all. We measure both *production probabilities* for different evidential devices under varying evidential circumstances, as well as *interpretation probabilities*, inferences about *speaker commitment*, and inferences about *the evidential circumstances* that generated the speaker's utterance. That is, we measure both sides of the communicative coin. In the process, we indicate how these case studies can profitably be linked to issues and controversies found in the current theoretical literature."

We would like to emphasize that acceptability ratings, which the reviewer proposes as an alternative dependent measure, suffer from arguably greater linking hypothesis problems than the more direct measures of production and interpretation probabilities we elicit here. Acceptability ratings are some function of grammaticality, semantic felicity, pragmatic felicity, and potentially other factors, none of which are clearly defined notions, and none of which can be independently estimated in such paradigms. In contrast, the measures we use here are more direct measures of production and comprehension.

[8] "What is the merit of using a quantitative measure over a qualitative measure (evidence types), apart from enabling or facilitating statistical analysis? Why are the data not presented in qualitative terms?"
Response: ***We have now added analyses of evidence type in addition to evidence strength to all sections of the paper. The original motivation for using evidence strength instead of evidence type was not to make statistical analysis easier, but to test whether evidence strength contributes independently to the choice of evidential device. If it does, this would require theories of evidential devices to be extended to include a notion of evidence strength that is not reducible to evidence type, despite the correlation. We agree with the reviewer, though, that it is pointless to not include the evidence type analysis if this point is to be made. Therefore, these analyses are now included.***

[9] "[…] This is particularly evident in the visualization in Fig.1: In place of the relatively unimportant distinction between the different lexical templates, I would have expected to see qualitative information on different evidential contexts. I would suggest that this information be incorporated in Fig.1. This way, one could immediately evaluate von Fintel & Gillies' proposal that *must* is only licensed in indirect evidential contexts."
Response: ***We have followed the reviewer's suggestion and included qualitative information about the evidential contexts in Figure 1. We also refer the reviewer to the Appendix, where the full set of stimuli is now listed.***

[10] "In their section 4, vF&G argue that indirectness is not directly tied to weakness, and that the two options should be strictly kept apart. They discuss a number of cases, in which utterances of *must p* come out with strong commitment (eg. ex. (4) is quoted from them), but these cases were not included in the current experimental paradigm. *They could be, though, in which case the current methodology might indeed make an important theoretical contribution in showing that indirect evidence and commitment strength are indeed independent phenomena (or not).*"
Response: ***We agree that this could be a very interesting point for our work. However, since this suggestion by the reviewer would require running a new series of experiments, we leave it to future (follow-up) studies to further address this point.***

[11] "A crucial factor not controlled for in the exps. is the factor strength of the inferential reasoning, which here does not involve logical deduction (indirect evidence), but a defeasible deduction based on stereotypical actions or chains of events. It is not surprising, then, that the commitment with *must/müssen* comes out as weaker. *This fact should be explicitly acknowledged in the discussion of the experimental results*."
Response: ***We have ultimately decided not to include this discussion, for the following reason: the direct/indirect distinction is itself not a very clear one, and that in turn means that logical deduction likely has a relatively minor role to play in the production and interpretation of evidential devices. For instance, in our items, the direct evidence for "The neighbor's dog is barking" is "You look outside and see Fluffy, the neighbor's dog, standing on the porch and barking", which was rated as weaker evidence than the following indirect evidence for "The coffee is cold": "You see that there is no steam coming from the coffee." (see Figure 1).***

[12] "The reported parallels in participants' behavior with epistemic *müssen* and the discourse particle *wohl* are not unproblematic, as – to the naïve reader – they might suggest a semantic parallel where none exists. As shown in previous literature, *müssen* and *wohl* differ in at least two important ways: (i.) *müssen* is truth-functional, can be questioned and be part of the presupposed background

etc; whereas *wohl* is not; (ii.) *wohl* is obligatorily not-at issue, for which reason it cannot be focused. So what can we learn from the fact that *müssen/wohl* behave on a par in the presented experiments? In all likelihood only that such experiments are insensitive to semantic differences between the two expressions."

Response: ***We slightly disagree with the reviewer, given that there is an important strand of work arguing for non-propositional analyses of modal expressions like (epistemic)*** **must** ***(see Groenendijk, Stokhof & Veltman 1996 for seminal work and Portner 2009 for a comprehensive overview of different approaches). Accordingly, the clear distinction suggested by the reviewer can also be considered controversial. Anyway, we now make very clear that the similar behavior of*** **müssen/wohl** ***in our study has no bearing on the debates indicated by the reviewer.***

"In other words, when investigating the dependence on evidence strength for *p*, we find that *muss* and *wohl* pattern together, in contrast to other modal means such as *vermutlich*. This is an interesting result given what we have discussed in Section 1.2 above, since our results suggest a use-conditional difference between discourse particles and otherwise synonymous adverbs in the domain of speaker commitment that has not been observed in the theoretical literature and that is due to use conditions (i.e., in which evidential environments to use these devices) rather than to fundamental semantic differences. On the other hand, we see that epistemic *müssen* patterns with *wohl* with regard to felicitous evidential environments, and this parallel again indicates similar use restrictions rather than semantic differences and/or similarities between modal expressions and discourse particles that are discussed in the literature."

[13] "Unlike *müssen*, *wohl* seems illicit in strong logical deduction contexts, such as (4), and the other strong contexts for *müssen/must* in vF&G (their exs. (13) and (14). Given this, the parallel between *müssen* and *wohl* in the experiments would follow from the particular (stereotypical reasoning) contexts employed in the experiments. When considering additional contexts, it might emerge that *wohl* is inherently weak (as has been argued in the literature), whereas the commitment strength of *müssen/must* p varies with inference scheme (and not with evidential context). *I would like to encourage the authors to engage in this kind of investigation, which may be more fruitful from a theoretical perspective.*"

Response: ***We agree that this is an interesting point/line of thought. However, in our study we aimed at shedding some light on the use conditions (in our case: evidential use restrictions) of these lexical devices. To test the hypothesis indicated by the reviewer, we would have to conduct an extra investigation where we carefully control for (more) deduction and inference schemes (see also response [14] below). We thus leave it to future work to address this issue.***

[14] "Whereas I agree that some of the experimental paradigms *might* be useful for shedding light on the theoretical debate, this would require controlling for the deduction or inferencing scheme, in addition to evidential contexts. *All this would need to be carefully reflected in the discussion section – which it isn't in the current version. For this reason, I see the imminent danger that theoretically untrained readers will draw too strong a conclusion from the reported findings! This must be avoided at all costs by inserting more hedging in the presentation and discussion of the experimental data.*"

Response: ***We revised the discussion section as follows, thereby introducing our conclusions in a more modest way now:***

"Although our experimental data cannot decide between the various theoretical accounts on challenging phenomena such as epistemic must (Section 1.1) and discourse particles (Section 1.2), a complete picture of these evidential means of natural languages requires systematically testing the compatibility of these devices with different degrees of evidence strength and evidential

circumstances; we have presented an experimental paradigm for doing just that. This paradigm will allow for the exploration of how different evidential devices are used and comprehended compared to alternative lexical choices. It also highlights the role of speaker commitment in theories of evidential devices. Below, we comment on some of the conclusions we can draw from our results."

In addition, we now make very clear in the final conclusion what the potential contributions of our experimental paradigm (which does not control for deduction or inference schemes, but instead focuses on evidential contexts) are:
"As for discourse particles, no study has to date taken into account the component of different degrees of speaker commitment when debating how, for instance, closely related sentence adverbs can be distinguished from synonymous discourse particles at the level of semantics. This is a new data point that needs to be accounted for in our theoretical understanding of the lexical inventory of evidential words in the German language.
As for epistemic *must*, we could, as expected, confirm the claim that the epistemic use of *must* expresses a weaker commitment than the bare form. However, in addition to experimentally confirming this (admittedly) trivial observation, we have tested for different degrees of evidence strength and also compared epistemic *must* not only to the bare form but also to alternative modal expressions available in the English lexicon (i.e., adverbs and *might*). In this context, we found that *must* is used in evidential circumstances where speaker commitment can be considered rather strong. This appears to concur with claims in the literature that highlight this strong component of *must* (e.g., von Fintel & Gillies 2010, 2016). However, 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* Φ are just as strongly committed to the prejacent as those who assert Φ by itself" (von Fintel & Gillies 2010: 30)."

[15] "Abstract: 'we provide a starting point for approaching theoretical debates on the nature of evidential expressions from an experimental and context-oriented perspective'
⇒ insert '(modal)' before evidential, as epistemic auxiliaries are not only evidential"
Response: ***Done.***

[16] "p. 5: 'Most of these differences are based on the assumption that *wohl* is not part of the truth-conditional content of the clause, whereas the evidential component of *must* and also the epistemic contribution by adverbs contribute to truth-conditional content. However, this distinction cannot be taken for granted, given what we said above about epistemic *must* and given the evidence for non-truth-functional views on higher adverbs discussed by Ernst (2007) and many others.'
⇒ I think the second clause in the quote is based on a misunderstanding: the semantic characterization of 'wohl' as not truth-functional does not hinge on its status as adverbial or not, but only on its observable behavior under negation and wrt presuppositions. The pointer to Ernst is superfluous as nobody is denying that there ARE non-truth functional adverbs, but the ones that compare to 'wohl' are NOT – as is obvious from their semantic behavior in questions and under focus."
Response: ***We thank the reviewer for this remark and added the following passage to illustrate the semantic behavior of*** **wohl** ***in questions (and how it differs from other modal devices like must and adverbs):***
"There are different views on how *wohl* might differ from adverbs and epistemic *must* concerning scope-taking in question formation and structured propositions (e.g., Zimmermann 2008, 2011). Most of these differences are based on the assumption that *wohl* is not part of the truth-conditional content of the clause, whereas modals like *must* and also the epistemic contribution by adverbs

such as *probably* contribute to truth-conditional content. Consider, for instance, the differences in the context of scope-taking in question formation. The formal sketches in (9), taken from Zimmermann (2011: 2021), make clear that using *must* in a question results in asking whether or not Max must necessarily be at sea, and the occurrence of *probably* yields a question that asks whether or not one has reason to suspect that Max is at sea:

(9) a. ⟦Must Max be at sea?⟧ = ?{Max must be at sea, ¬(Max must be at sea)}
b. ⟦Is Max probably at sea?⟧ = ?{ASSUME(x, Max at sea), ¬ASSUME(x, Max at sea)}

The point is that in both cases, the semantics of these modal devices forms part of the alternatives under discussion, according to Zimmermann (2011). The modals hence contribute to the propositional (and thus also to the truth-conditional) content of the utterance. With this observation in mind, let us now look at one example given by Zimmermann (2011: 2020) that exemplifies the scope-taking behavior of the epistemic particle *wohl* in questions:

(10) a. Hat Hans wohl Maria eingeladen?
has Hans PRT Mary invited
'What do you reckon: Has Hans invited Mary?'
b. WANT (S, A, know (S & A, ASSUME {Hans invited Mary, Hans did not invite Mary}))

Both the translation (10a) and the formal sketch by Zimmermann (2011) in (10b) indicate that the semantics of *wohl* does not form part of the alternative propositions, in contrast to what we have seen in (9) for other modal devices. In particular, the question in (10a) is not asking whether or not there is a lack of commitment towards the proposition. Rather, by using *wohl*, the speaker wants the addressee to make their best guess concerning the alternative answers, and this can be expressed by the operator ASSUME, which takes scope over the alternative answers rather than being part of them."

However, we also would like to keep our claim that differences between* must*/adverbs and* wohl *are more controversial at the level of semantics than at the level of syntax. First, the observable behavior under negation is the same for adverbs like* vermutlich *(see our example (7)). Second, we would like to point out that Grosz (2016), Cardinaletti (2011), and many others argue for an adverb theory of discourse particles. And third, please note that* wohl *in questions (the main data point discussed by Zimmermann 2011 above to show that* wohl *cannot be compared to the corresponding adverb) differs semantically from* wohl *in declaratives (our data point). One might even say that they are actually two different lexical items/particles, as is the case for many other particle examples in German like* nur *in questions vs. imperatives,* ja *in declaratives vs. imperatives, etc. Therefore, we revised the relevant passage as follows, also making clear that these debates have no bearing on our context-oriented experimental investigation:
"[…] Be that as it may, recall at this point that there is considerable debate in the literature on which semantic dimension is the appropriate level to account for the evidential component in the meaning of *must*. In other words, there are many approaches according to which the epistemic reading of *must* is not a truth-conditional part of utterances (see references in Section 1.1 above). Also, given the evidence for non-truth-functional views on higher adverbs discussed by Ernst (2007) and many others, one may conclude that the differences between epistemic discourse particles and other modal devices for expressing analogous meanings can clearly be seen at the level of syntax (see (7) and (8) above), while the differences are more controversial at the level of semantics (see also Grosz 2016 on this point).

In what follows, we will abstract away from these core-semantics issues and instead add a new aspect to these debates, thereby addressing a blind spot in the theoretical literature. That is, we will approach the different modal devices mentioned above from a pragmatic (i.e., context-oriented) perspective by investigating the extent to which *wohl* differs from its modal counterparts in its compatibility with different evidential contexts. Although all evidential devices (epistemic *must*, adverbs, and discourse particles) modify the speaker's commitment to a proposition, no study to date has explored whether these (at first sight synonymous) expressions differ in their compatibility with different evidential circumstances."

[17] "p.6: 'Thus, in this section we report on an experiment that collected estimates of evidence strength with the goal of norming evidence types for a variety of characteristics that will serve as the basis for the studies to be presented below.'
⇒ Once again, what is the advantage of a quantitative measure of evidence strength over a qualitative definition of evidence types?"
Response: ***We now report results for both quantitative and qualitative characterizations of evidential contexts; see above.***

[18] "p.6, below (9): 'evaluated one of five possible pieces of evidence, resulting in four trials per participant'
⇒ These five qualitatively different pieces of evidence should be explicitly listed. This information is much more important than the different lexical frames in (9)! Same in Fig.1!"
Response: ***Figures 6 and 7 now report results for both quantitative and qualitative characterizations of evidential contexts.***

[19] "p.6, 2.2: 'We obtained between 3 and 14 strength ratings for each piece of evidence.'
⇒ Why this huge variance in the number of ratings?"
Response: ***Pieces of evidence were assigned at random to our participants, which accounts for the discrepancy in the numbers of ratings obtained.***

[20] "p.7, Fig1: Figure 1 is difficult to interpret: what subtypes of evidence/contexts trigger strong ratings, which ones do not? The sorting by lexical material is of secondary importance!"
Response: ***We have included information about evidence type (i.e., a qualitative characterization) in Figure 1.***

[21] "p.8, 3.2: 'The overall distribution of utterance choices is shown in Figure 2.'
⇒ Is this information relevant? Perhaps leave out, as the proportions also hinge on random the selection of 3/5 contexts, which might induce an artificial bias."
Response: ***We have opted to keep this figure in in order to demonstrate the baseline utterance preferences in our experiment. As the reviewer notes, this baseline does depend on the random selection of contexts. For this reason, we have included bootstrapped 95% confidence intervals indicate which differences are reliable.***

[22] "p.10, end of §3: 'In other words, when investigating the dependence on evidence strength for p, we find that *muss* and *wohl* pattern together'
⇒ This statement makes reference to Fig.4, right? There is no discussion of the statistics behind 4 - as far as I can see –, and it is not entirely clear from the visualization that *wohl* behaves on a par with *müssen*. Perhaps it would be better to replace the proportions of 'wohl' and 'vermutlich', such

that the weakest ‘vermutlich’ comes topmost in parallel to English ‘might’, thereby giving the visualizations an overall more similar appearance.”
Response: ***This statement in an elaboration of the sentence that precedes it, which makes explicit mention of Fig. 3. As Fig. 3 shows, there is no discernable difference between*** **muss** ***and*** **wohl** ***in terms of evidence strength. As for Fig. 4, we now mention it explicitly in the prose.***

[23] “p.12, middle: ‘Fig.6 shows that...’
⇒ Please add a pointer that Fig.6 is found only much later in section 5!”
Response: ***Done.***