**Overview of reviewers’ evaluation**

We thank the reviewers for their highly constructive comments on our previous submission. The reviewers’ guidance was critical in revising the manuscript. To briefly recap, all three reviewers seem to agree that this is in many ways a very strong manuscript that makes substantial methodological (R1-R3), empirical (R1-R3), computational (R1, R3), and theoretical contributions (R1). However, all reviewers also raised concerns about our—poorly worded—claims to novelty, and—as R3 put it—the amount of material one has to “wade” through to come to the most novel contributions. We believe that these two inter-related concerns were primarily rooted in presentational issues, and the reviewers’ comments were very helpful in addressing those.

Specifically, R1 describes the manuscript as a “very strong paper, both empirically and theoretically”, highlighting that “it addresses long-standing gaps in our theoretical understanding of the fine-grained timecourse of listeners adapt to a novel talker and precise details about the mechanisms …”. R2 agrees that “there is a great deal of thought-provoking material in this paper”, “(ultimately) some valuable implications emerge”, and that our data analysis is “thorough” but had concerns about some aspects of our “stimuli, the placement of the work in prior literature, and the novelty of the paradigm”. R3 described the work as “sophisticated and methodologically well done”, with “several strong methodological innovations including the use of multiple testing points, the psychometric analysis […] and the ideal observer model”. While highlighting that the paper is dense and seems to largely confirm known findings, R3 concludes that our finding of ‘premature convergence’ is “super novel and super interesting” and “the most compelling aspect of the paper”.

**Overview of revisions**

The reviewers’ critiques were eye-opening. On re-reading our manuscript with the reviewers’ feedback in mind, we realized how little we had motivated our approach, and, in particular, why we agree with a number of recent reviews that *much less than commonly assumed* is known about the mechanisms underlying rapid adaptation. This specifically includes that it is *not* known to what extent distributional learning can explain the surprisingly rapid changes during speech perception that often emerge within less than a handful of trials—at time scales that might differ starkly from L2 learning (which R2 focused on) and many other learning phenomena beyond speech perception (which R3 focused on). Of course, any researcher might have their own intuitions about what is and what is not a plausible explanation of rapid adaptive changes in speech perception. However, as the revised introduction now clarifies, there are large parts of the field that do *not* consider distributional learning the most likely candidate hypothesis. And that is, we suspect, in parts due to the rather weak nature of the existing tests (our own previous work included).

Before we outline our revisions, we briefly reflect on the notion of “strength of evidence” or the “strength of a test”, which plays an important role in the revised introduction—the idea that there is a difference between testing a binary (A > B) hypothesis vs. the type of fine-grained evaluation that assesses whether a model or theory can actually explain both qualitative and quantitative aspects of the data (in our case, changes in speech perception as a function of the phonetic distributions in the previous and current exposure). Consider the following hypothesis tests:

1. Whether exposure affects subsequent perception relative to some other exposure conditions, without any prediction about the direction of the effect (e.g., accuracy after exposure A != accuracy after exposure B)
2. Same as 1 but with a clear expectation about the direction of the effect (e.g., accuracy after exposure A > accuracy of after exposure B).
3. Same as in 2 but while also making predictions specifically about how the phonetic properties of exposure predict the changes in the mapping from phonetic features (e.g., VOT) to phonetic category (e.g., /d/ vs. /t/). For instance, exposure A uses VOT distribution that is right-shifted relative to exposure B; following exposure, listeners in condition B categorize fewer token along the VOT continuum as “t”, compared to participants in condition B.
4. Same as in 3, but for more conditions and measured at more points time, so that the directional prediction for conditions A, B, … are tested repeatedly, and the timecourse of those changes becomes clear.
5. Same as in 4, but while also having a predictive model of *exactly how* participants’ categorization function across the entire VOT continuum should change with the type and amount of exposure.
6. Same as in 5, but while also testing whether the direction and magnitude of the changes are compatible with a model of participants’ *prior* expectations *based estimates of the relevant phonetic distributions in their previously experienced speech input.*
7. *Etc.*

Most existing work on adaptive speech perception presents tests like those in bullet 2 or 3 (qualitatively contrasting two exposure conditions, or qualitatively showing that listeners with different prior experience perceive the same stimulus differently). These types of studies provided important proofs of concepts that distributional learning is broadly compatible with the available data. They do, however, provide rather weak tests of the hypothesis that distributional learning is thecore mechanism underlying rapid changes in speech perception (a hypothesis that has been explicitly called into question). As others have pointed out long before us (e.g., Newell; Platt; Yarkoni; etc.), there are times in a field when binary tests like those in 2 and 3 above—what Newell calls the “game of 20 questions”—cannot forcefully advance theory.

*No study we know of* (including those mentioned by the reviewers) presents tests of type 5-6—i.e., tests that actually assess to what extent changes in listeners’ perception *are qualitatively and quantitatively explained by the phonetic distributions in the relevant input*. Such tests are, however, crucial. Even if one—on contrast to several recent reviews of the field—were to take for granted that adaptive speech perception relies on distributional learning, the existing evidence would leave open whether distributional learning can actually explain a non-trivial share of participants’ adaptive behavior—as would be expected if distributional learning is the core mechanism underlying adaptive speech perception.

The study we present takes an important step towards addressing this question. This is why we evaluate changes in listeners’ behavior for a total of >400 combinations of preceding exposure + current test stimulus—each of which is predicted to have a unique p(“t”-response | phonetic distributions of /d/ and /t/o prior to experiment, phonetic distributions of /d/ and /t/ in experiment experienced so far). As the revised introduction now clarifies, we compare these measures against both idealized listener/learner models, and an actual distributional learning model. It is precisely this model-guided ‘stress test’ that leads us to make several important empirical contributions:

* As R3 notes: “frankly the sophistication of the ideal observer models […] makes it starkly clear that the subjects are [not converging against the statistics of the input] by providing a very clear view of what the subjects ‘should’ be doing.” Critically, if we had analyzed our data following standards in the field, our results would have appeared to follow the predictions of distributional learning theories—as they did in the many studies the reviewers cited.
* The same careful, model-guided approach could also have also disconfirmed some of the other predictions of distributional learning. For instance, we know of no other study that predicts the direction of the change from pre-test to post-test. This might sound trivial, but it requires at least something akin to our idealized models of pre-exposure listeners and post-exposure learners. That our approach did not find the other predictions disconfirmed thus is an important finding, too, though we recognize that it might appear perhaps less ‘surprising’ than the finding R3 highlights.
* Overall, a model of distributional learning derived from general principles of information integration (the ideal adaptor) provides a *very good* quantitative model of adaptive speech perception (R2 > 96%) for our stimuli and paradigm, while also failing to predict one fundamental constraint on rapid adaptive speech perception. To the best of our knowledge, this is the first evaluation of an actual distributional learning model of this type against incremental adaptation to distributional input (we discuss related work in the introduction). The model is constrained to have prior beliefs that have to match the phonetic distributions experienced by a ‘typical’ listener (as estimated from a phonetic database of US English), and only has two parameters that determine how these prior beliefs are updated to fit listeners’ rapid incremental adaptation to shifted distributions in the phonetic input. It would have been entirely possible that it captures only the general qualitative trend of human behavior, rather than such a high share of the variance.

We hope that the revised framing highlights why we are excited about, not only the findings, but also the model-driven approach we present. R2 and R3 are absolutely correct: it’s not the *paradigm* by itself that is novel (contrary to what we originally wrote); it’s the combination of that paradigm with the idealized listener, idealized learner, and ideal adaptor models that, we believe, (1) sets new standards for research on rapid adaptation in speech perception and beyond and (2) results in a thought-provoking novel finding, the theoretical consequences of which we continue to review in the general discussion.

**What has changed?**

* Title
  + We changed the title to more clearly signal the goals of the paper. It now reads: “Learning to understand an unfamiliar talker: Testing distributional learning as a model of rapid adaptation during speech perception”
* Introduction
  + We clarify early on that we are focused on *rapid* changes in speech perception at time scales very different from L2 learning (addressing concerns of R2). This, of course, does not mean that L2 learning work is irrelevant to our goal—it is *possible*, for instance, that the same mechanisms underly both rapid adaptation and L2 learning.
  + We motivate our approach by long-standing criticisms of how research in the cognitive sciences tends to proceed through a “game of 20 questions” (Newell). This motivates our goal to take a more holistic approach, developing stronger, quantitative evaluation (Westfall & Yarkoni, 2017) of how distributional learning as a theory of adaptive speech perception—a contribution beyond just research on adaptive speech perception.
  + For presentational reasons, we switched the number of predictions 3 and 4. Prediction 3 now refers to the *diminishing returns/power law of learning*. Prediction 4 refers to *premature convergence*.
  + We anticipate the unexpected finding (‘premature convergence’) and its potential explanation in the abstract & introduction (following R1’s advice). We also clarify early on that this is indeed the prediction we are most interested in (as surmised by R3).
* Methods:
  + We motivate our design decisions in the Methods section, linking back to the predictions and overarching goals described in the introduction (following advice from R1 & R3).
  + We introduce the idealized listener and learner models in the Methods section, so that they are common ground before we discuss the results (following R3)
  + We renamed the exposure conditions (following R2’s advice) and clarified in more places that exposure was manipulated *between* participants (addressing R3’s questions).
* Results:
  + We moved the comparison of the ideal adaptor (distributional learning) model from the general discussion into an early result section. This new section now explains the model in more detail, and presents new tests that confirm the premature convergence result (requested by R3). We hope this helps to clarify *how* the 400+ measurements we collect help to identify previously unknown constraints on distributional learning.
  + The new structure of the result section—the fact that we first present general data analysis (through our Bayesian mixed-effects psychometric models), followed by additional computational modeling is anticipated at the start of the Results section.
  + To keep the manuscript from becoming too long, we moved both the conceptual replication of previous work and the analysis of repeated testing to the SI (the latter is, however, briefly summarized in the main text, as it is important in understanding some of the other results).
* General discussion:
  + We have restructured the general discussion to more explicitly discuss the different contributions of our paper. This includes a new section that briefly discusses the merits of quantitative tests of actual *models*.
  + We now also anticipate that the discussion focuses on the most surprising result (failure to converge against distributions in the input). The section that discusses this finding has been moved up, and has been expanded to integrate feedback from reviewers.

Throughout the paper, we have integrated reviewers’ suggestions, including additional literature, corrected/adjusted our claims, and provided more guidance for readers. The additional clarification required additional space, and we had to make tough decisions as to what to introduce where in the paper. We hope that this can be appreciated, and that the revised manuscript makes clear how this study advances the field methodologically, empirically, and theoretically. We again thank the reviewers, and look forward to additional suggestions they might have to improve the paper.

Next, we provide point-by-point reply to reviewers’ feedback. **Our responses are in green.**

**Reviewer #1**

[summary omitted]

Overall, this is a very strong paper, both empirically and theoretically. It addresses long-standing gaps in our theoretical understanding of the fine-grained timecourse of listeners adapt to a novel talker and precise details about the mechanisms for the adoption of a representation of that talker's shifted speech patterns. The analytical approach in the paper is cutting edge and very clear - I also personally really liked how the results section was divided in sections labeled according to which specific research question each sub-analysis addressed. Even the table captions provide the specific research question addressed in the statistical outputs provided for each - really nice touch with that. I have only minor comments for the authors.

Thank you! We very much appreciate the encouragement!

Points for the authors to consider:

1. First, this is a highly theoretical paper about how and when adaptation occurs - which is great and will be an important contribution to the literature. But, the empirical observations are based only on an initial stop voicing contrast (even, just 4 /t/-/d/ sets). I think the sweeping and broad claims about the ways in which this study demonstrates how adaptation works should be tempered throughout given that the phonetic features (and items) examined in this study are quite limited. For instance, Do the authors predict the same time course/nature of adaptation for novel vowel shifts?

We very much agree. We had mentioned this limitation in the Limitations sections. We now elaborate on this point—including the reviewer’s point about different phonetic features (no, we would not expect the same time course for different features, e.g., spectral vs. temporal features). We now also repeat this caveat in the conclusion section. Additionally, we have revised a few places in the result sections and in the general discussion to remind readers (and ourselves) that our findings are observed for a particular set of stimuli and a particular task.

2. In the abstract, the experimental paradigm is described as completely novel. While it is true that the precise format of this study has not been performed previously, it is a bit of an overstatement to claim this is a completely novel experiment (aspects of the study are conducted across many different adaptation studies). The paper is very strong without this claim - I recommend not overstating the novelty of the experimental design.

Point taken! This is something also raised by R2. We therefore have addressed the point in the letter to the editor.

3. In my opinion, the description/motivation/discussion of prediction 3 was less clear than those for 1, 2a/b, and 4. In the discussion of prediction 3, some interesting alternatives to distributional learning are raised that I think do provide an explanation for the "premature convergence" observed - namely, that it is possible that listeners have previously encountered talkers like the one they are being exposed to and thus they already have a representational model to employ when given even the briefest amount of evidence from this talker. I think that is really what is going on. In fact, I believe there is some recent work in the sociolinguistics literature that supports this idea, too. For instance, see the work of Lacey Wade *(Wade, L. (2022).* *Experimental evidence for expectation-driven linguistic convergence. Language, 98(1), 63-97)* showing that when listeners are given a small amount of evidence that a talker might have a southern American English speech system based on just one phonetic feature, they show a shift in another feature that is also typical of SAE even if they had no exposure to that shift. Studies like this also are evidence that listeners might be trying to select the right speech model for a given talker early and with, in some cases, incomplete information. This is indeed discussed substantively in the discussion section. But, I recommend the authors bring up these such possibilities/alternative to DL in the introduction to help guide the reader to expect what is actually observed - and, again, it is not surprising given some other recent works.

Thank you. We followed the reviewer’s suggestion. The revised introduction now anticipates the point. In this context, we now also cite Wade (2022). In the general discussion, we further elaborate on the link to recent sociolinguistic research that the reviewer kindly made us aware of.

4. Some of the in-text citations have author initials (e.g. AA or M.M.)

5. Page 9: "we find little support for prediction (3 - learn to convergence)." It is odd to have the 3 in the parenthetical, no?. Should this be: we find little support for prediction 3 (learn to convergence). Also, on the subsequent pages - it is odd to have the number in the parenthetical.

6. Figure 6: the labels are overlapping in the 3 lower right corner panels - please fix

We fixed all of these points except that (6) since that information is not critical (and due to the fact that the means are very similar anyway, which we believe is visually clear). Thank you for catching these issues!

**Reviewer #2**

[summary omitted]

There is a great deal of thought-provoking material in this paper. The argument is complex but ultimately some valuable implications emerge. The analysis uses an appropriate analysis method (Bayesian mixed-effects psychometric models) and does so very thoroughly. There are however weaknesses with respect to the stimuli, the placement of the work in prior literature, and the novelty of the paradigm.

We thank the reviewer for the balanced assessment. As described in the main part of this letter, we have completely revised the introduction to clarify the contributions of our paper, which—as the reviewer correctly points out below—do not lie in the novelty of the paradigm, but in the strength of the tests afforded by the novel combination of paradigm, analyses, and model-guided interpretation.

1. Stimuli. As I was reading the Methods section, I found myself looking for (and failing to find) justification for the choice of stimuli across the three conditions. I had three questions: Why means of 5 and 50 ms in one condition, why was this a "baseline", and why +10 and +40 ms for the other two conditions? I was also surprised to read the last sentence of section 2.3 ("We note that the naming of conditions (baseline, +10, +40) should be understood as relative to each other, rather than relative to listeners' prior experience"). I wanted to know how these distributions lined up relative to the participants' prior experience (especially given Prediction 1, as set up by the authors in the Introduction). I was eventually given answers to this last question in the middle of the Results section (Figure 6 and accompanying text). But there were still no direct answers to the first three questions. The indirect/implicit answers were that the baseline means were not selected in a way that justified them to be called a "baseline" and that the shifts of the means in the other two conditions were not motivated on the grounds of estimates of prior distributional knowledge.

Correct. We apologize for the confusion our wording might have caused. As laid out in the main part of this letter, we have revised the paper to be clearer that the naming of the conditions was essentially arbitrary. We could have named the conditions A, B, & C. What matters is that these conditions are shifted relative to each other, and relative to listeners’ prior expectations. Our intentions for naming the first condition “baseline”, along with the statement the reviewer quotes, was precisely to alert readers to this arbitrariness. We do now, however, follow the naming suggestion the reviewer provides below.

I think that the optimal solution to this issue would be to rerun the experiment, potentially with three conditions: a baseline with means matching the means from the Chodroff and Wilson dataset, and two other conditions, one with a shift to VOTs lower than these baseline means and one with a shift to higher VOTs. Instead of seemingly arbitrary, fixed, and matched absolute shifts (i.e. +10 and +40 ms for both categories), it might be wise to base the shifts on the Chodroff and Wilson data (e.g. +/-1.5 SD, thus ensuring a larger shift for /t/ than for /d/, in line with the greater naturalistic variability in /t/). I predict that such a design would result in larger and clearer effects of prior knowledge and experiment-internal exposure and (hence) also that more subtle effects (how quickly does learning take place, when does it plateau) would be clearer too.

There are many possible exposure scenarios that could, and should, be compared in future research. However, it is not clear which ones of them would be more informative—in part for all the reasons we now lay out in the revised introduction: while a lot is known about the *qualitative* effects of exposure, very little is known about the *quantitative* effects of exposure. But without actual *models* that make such quantitative predictions, there is no objective criterion that makes one exposure condition ‘better’ (or more informative) than another.

For example, while the reviewer’s prediction about their preferred design strikes us as plausible, it’s perhaps based on intuition or experience with previous experiments, rather than an existing model? As the revised introduction now clarifies, an important motivation for the present work is to move beyond such intuitions *because intuitions can be misleading, and violations of intuitions are not necessarily informative for theory development.*

2. While I think this new experiment would make for a better paper, I don't think the current experiment is unpublishable without it and so I will not insist that this new experiment be added. But I do think the current paper needs to be substantially revised, in three ways. First, I think the Chodroff and Wilson data should be presented much earlier, either in the Methods section or perhaps best in the Introduction. Second, stimulus selection needs to be motivated relative to these data. Maybe there is a motivation for why the current "baseline" condition deserves the special status of "baseline" that I am missing, but if not (and especially if there is also no strong motivation for +10 and +40) I would suggest (in line with my suggested experiment above) that the current +10 condition be referred to as the "baseline" and the other two as -10 and +30. The reason for this is that, as shown in Figure 6, the means for the current +10 condition are both not at the tails of the Chodroff and Wilson distributions, while the /d/-means for baseline and +40 are at the left and right tails, respectively, of the /d/ distribution. There is then at least the motivation, with respect to /d/, that the current "baseline" has an extremely low mean, and the current "+40" has an extremely high mean. This motivation doesn't work for /t/, but it might be enough to help clarify to the reader (earlier than the results section) how the conditions relate to prior knowledge. Third, if these changes are made, the entire results section would need to be re-done, with new condition labels. While these changes may appear superficial (the results themselves won't change), I believe there would be substantial gains in clarification of what the results are and what they mean with better up-front motivation of the conditions relative to (the best estimate of) prior knowledge.

We agree, and have more or less followed the condition naming suggestion of the reviewer. We believe it has made the paper more accessible. Thank you!

We note that the naming of conditions is based on the predicted PSE relative to prior experience (rather than the means of the /d/ and /t/ category relative to prior experience), since this is the measure we use to compare human behavior against the distributional learning model.

3. Prior literature. Even though a broad array of prior studies is discussed, I felt that there was insufficient acknowledgement of the research on which the current study is built, and that these acknowledgements needed to be made in the Introduction to set the experiment up. First, and more generally, I think there needs to be more discussion of the work that has been done on distributional learning underlying the acquisition of novel sound categories in non-native languages. Although this work is mentioned (e.g. in Footnote 2), this brief discussion does not do sufficient justice to the literature which has explored issues that are addressed in the current work (e.g. Best's PAM addresses the powerful constraints that prior knowledge about the distribution of phonetic cues in L1 can have on learning about L2 categories; see e.g. Escudero et al., 2011, on effects of the experiment-internal distribution). Second, and more specifically, the ways in which the questions addressed in the current study are introduced appear incomplete. In the lexically-guided perceptual learning literature, several studies on the amount of exposure required for learning to take place are not cited (Kraljic & Samuel, 2007; Poellmann et al., 2011). Similarly, earlier work on the effects of prior knowledge in accent learning (e.g., Witteman et al., 2013) and the effects of exposure distributions in learning L1 categories (e.g. Zhang and Holt, 2018) should be acknowledged. The bottom line is that, while the earlier work may have manipulations that may not always be as fine grained as those that are tested here, nor are they exactly the same manipulations, that earlier work should nevertheless be acknowledged as providing motivation and context for the current study.

Thank you for the pointers to these additional papers (some, but not all of which, we were aware of). We have integrated some of these papers into the introduction where appropriate.

We note that the studies referenced by the reviewer present qualitative tests of the effects of prior knowledge and exposure distributions. None of these studies tests to what extent a model of distributional learning can explain the results—i.e., whether the differences in listeners’ behavior follow from the phonetic distributions in their input. Of course, these are important studies in their own right. But, as we now clarify in the revised introduction, they differ in important ways from what we aimed to achieve in the present work.

4. Novelty of paradigm. A related point is that the exposure-test paradigm used in the current study is not as novel as it is presented as being. There is a very large literature on L2 sound acquisition that has used (variants of) paradigms in which the effects of learning are tracked over time (e.g. with blocks of testing interleaved with blocks of exposure, and/or with measurement of learning performance over time in exposure trials, and in both cases with pretests and posttests). See, for example, the classic study by Logan, Lively, and Pisoni (1991) on L1 Japanese participants learning the L2 English /r/-/l/ contrast and the research inspired by that study. Although many studies on lexically-guided perceptual learning do not include pretests, some do (e.g. Eisner & McQueen, 2006). The classic Bertelson et al. (2003) study on visually-guided perceptual learning has interleaved exposure and test blocks. In short, I think it is incorrect to describe the experiment as having "a novel incremental exposure-test paradigm" (abstract); it is rather an adaptation/application of a well-established and widely-used paradigm.

We agree. This was unfortunately worded. We did not mean to suggest that the idea of incremental exposure and testing is itself novel. The revised manuscript does not present the paradigm itself as novel. Instead, our most important methodological innovation is the *combination* of the incremental exposure-test paradigm with model-guided data analysis/interpretation; it is this innovation that is allows us to make the novel contributions we present.

We hope the reviewer does not mind if we provide some context here. We are, of course, inspired by the seminal work by Logan et al. (as well as other early works from the McClelland lab) on distributional learning over speech inputs. But this and similar studies look at how adaptation unfolds over much longer periods of time (weeks!). This makes sense, of course, given that they focus on L2 acquisition. How the acquisition of L2 phonological categories unfolds over weeks of explicit training was—and is—a fascinating question. But it is a separate question how rapid adaptation over mere minutes of exposure affects L1 speech perception (and in the absence of any “training”: unlike in L2 learning studies, our participants were not asked to learn a new language; they were simply listening to someone speaking in listener’s L1). While it is quite possible that rapid changes in L1 perception originate in the same mechanisms as L2 acquisition, that is by no means taken for granted in the literature (see recent discussions in Zheng & Samuels, 2020; Baese-Berk, 2018; Bent & Baese-Berk, 2021; Xie et al., 2023).

For example, one alternative hypothesis frequently entertained in neuroimaging work is that rapid adaptation during speech perception is achieved by changes in decision-making (Myers & Mesite, 2014), rather than distributional learning. Other hypotheses in the literature vaguely refer to “criterion relaxation” or “threshold changes” as alternatives to distributional learning. To the best of our knowledge, the present study is indeed the first to assess distributional learning theories incrementally during the early moments of exposure to an unfamiliar talker.

This is also different from Bertelson et al or Vroomen et al. (2007) and related works. Like the work by Pisoni et al, these studies very much inspired the present work. (We continue to cite many of them in the introduction, and returned to them in detail in the general discussion). But all of these studies investigate the incremental exposure to the exact same labeled stimulus. And none of them actually investigates whether the phonetic properties of this stimulus explain the observed changes in listeners’ behavior.

Similar to our point about L2 acquisition, it is possible that adaptation in such repeated-stimulus paradigms draws on the same mechanisms that underlie adaptation to distributional exposure, and that this involves distributional learning (see discussions in Zheng & Samuel, 2020; Cummings & Theodore, 2023). However, as also pointed out by R3, we cannot simply assume this to be the case. For instance, a common criticism of distributional learning theories is that they raise unaddressed questions about the ability to maintain and integrate exposure information across time (an area where, e.g., exemplar theory and ideal adaptor theory differ in important ways). But these questions do not arise if the stimulus presented on each trial is identical. There are other important differences between these paradigms and ours that affect how likely participants’ behavior in the different paradigms is likely to generalize to everyday speech perception. We continue to discuss those differences in the general discussion.

Additional comments

The last sentence of the abstract would be more informative if it indicated what the "previously unrecognized limits on adaptivity" are.

Both the abstract and the introduction now anticipate the nature of the unexpected result (the ‘premature convergence’).

p. 10: Why 126 participants (42 per group, after exclusion approx. 40 per group)? Was a power analysis performed?

We did not perform a power analysis for this particular study. Standard power analyses would have been uninformative given that this is the first study of this type (for issues with the common practice of conducting power analyses over assumed effect sizes, such as the power to detect a “moderate” effect, see Xie et al., 2023). Instead, our sample sizes were based on several previous (non-incremental) distributional learning studies in our lab (citations omitted for the sake of anonymity).

p. 15, top: In the labelled trials, did one of the two response options correspond to the stimulus (e.g. stimulus "dill", response options "dill" and "din")? If so, then it isn't completely correct to say that there was no lexical disambiguation on these trials (p. 15). The stop is labelled because both words start with "d", but the word is also labelled and this (potentially) provides additional labelling of the word's component sounds (imagine of the response options were "Xill" and "Xin", with no /d/ label; the match of the stimulus to the word "dill" still gives, indirectly, the /d/ label). Or were the labels always for other words (e.g. stimulus "dill", response options "din" and "dip")?

The former: e.g., for stimuli along the *dill-till* continuum, a /d/-labeled trial might have response options "dill" and "din" (or “dill” and “dip”). We explicitly avoided introducing scenarios in which none of the response options matched their perception of the stimulus.

We did not quite understand this comment “If so, then it isn't completely correct to say that there was no lexical disambiguation on these trials (p. 15).” We never claimed otherwise. Yes, labeled stimuli were labeled lexically (or, rather, pragmatically: even if the listeners didn’t perceive a dill-till recording as “dill”, the absence of a “till” option would strongly bias the listener to infer that the input was intended to be “dill”).

Perhaps the reviewer is referring to this passage on p. 15:

*While lexical context often disambiguates and labels sounds in everyday speech …, disambiguating context is not always available. Especially with unfamiliar accents, listeners often have uncertainty about the word sequences they are hearing, reducing the labeling information available to them. Here, we thus struck a compromise between never or always labeling the input.*

This passage refers to the availability of lexical labeling in “everyday speech perception”, i.e., outside of experiments. The paragraph is meant to motivate why we used a mixture of both labeled and unlabeled exposure (though results from Kleinschmidt et al., 2015; Kleinschmidt, 2020 strongly suggest that fully labeled or fully unlabeled exposure would not have qualitatively changed the results in this type of paradigm). We have revised this paragraph to clarify this. If the reviewer has additional suggestion for rewording to make this clearer, we are happy to integrate them.

p. 23, footnote 7: This is confusing. Why use a test and then say it isn't appropriate?

Thank you for highlighting this. We have removed the footnote, as it was introducing unnecessary confusion (we were being overly cautious). For readers who want more information about the null effects, we now also include the “probability of direction” in our result tables. The probability of direction is an index of how much support there is for the presence of *any* (non-null) effect. It is the proportion of the posterior samples that have the same sign as the mean of the effect. If an effect is null, the probability of direction should be close to 50% indicating that half of the posterior falls within either side of 0. This is what we find for the very few (non-critical) null predictions.

Table 3: Explain in a table note why some rows are italicized. This currently has to be inferred from the main text.

p. 9, l. 209: Figure 2C -> Figure 2D

p. 29, l. 598 : Panel B -> Panel D

p. 30: Figure 7 is not discussed in the main text.

p. 32: Spell out VG and LG in VGPL and LGPL on first use of these abbreviations.

p. 40, l. 887: improve -> approach

We fixed all of these. Thank you for pointing them out!

p. 45, ll. 1008-1009: "the ideal adaptor substantially under-predicts changes in listeners' PSEs during initial exposure, and over-predicts changes in listeners' PSE following exposure". I do not see this - the error bars are overlapping with the ribbons. Am I missing something?

Given the importance of this analysis, we now describe it in more detail in a separate Results section before the general discussion. The reviewer is correct that the CIs overlap with the model predictions (which is directly related to the fact that the ideal adaptor achieves a high R2 of 96%).

However, even great models can be partially wrong =). We now present additional Bayesian hypothesis tests that assess the claim we made based on the figure. These tests find very strong support for our point. Specifically, there is a subtle but consistent, qualitative mismatch between the model’s predictions and listeners’ behavior: the model will always predict convergence with sufficient exposure, whereas listeners seem to plateau. In the figure the reviewer referred to, this shows as the model predicting less steep changes in the start of the experiment than observed, and predicting more steep changes at the end (it’s the best the model can do to fit listeners’ behavior).

**Reviewer #3**

One of the hottest areas in speech perception of the last decade or two has been the astonishing plasticity of the system. There has been a very large number of demonstrations that adult perceivers are highly plastic and can rapidly retune their perceptual systems to cope with new contexts. The present study offers a very carefully done contribution to this literature. They took a comprehensive look at the whole process, starting from a pre-test of the initial category structure, and then several rounds of "training" followed by test to see how the learning unfolds over time. This was done in a distributional learning framework, which is importantly different (and potentially more general) than the lexically guided retuning paradigm which has dominated most recent work. It is analyzed with very nice psychometric approach and the analysis is strengthened by the inclusion of a variety of ideal observer type models that help establish what to expect with input

The work is sophisticated and methodologically well done. There are several strong methodological innovations including the use of multiple testing points, the psychometric analysis is novel (and wholly appropriate) and the ideal observer models are very helpful. The basic learning paradigm - while borrowing heavily from others - is also interesting in its mix of unsupervised and softly-supervised trials. There's little to critique on methodological grounds. In some ways, this is the most comprehensive evaluation of a basic distributional learning paradigm that I have seen (I say basic because it doesn't look at any of the more interesting recent variants such as learning multiple talkers, comparing supervised vs. unsupervised, etc).

We thank the reviewer for this summary, and the encouragement.

At the same time, I was left with an unclear sense of what the basic contribution is. The authors start with several key questions: whether learning depends on prior distributions, whether it depend on the amount of exposure, whether there are diminishing returns, and whether learning fully includes the new "distributions" or stops prematuring. Of these, the first four seem fairly non-controversial. Just to briefly describe what I mean…

First, the point that learning depends on prior distributions. Well of course it does. This non-controversial. The authors kind of pitch it in a soft Bayesian approach, but all models of adaptation of the L1 assume that what is going on is that learners adapt their existing categories. To that end, the critical empirical novelty is that there's a pretest. But even then (as I detail below) it is not clear that the pretest really tests this hypothesis.

As we now clarify, it is not the pre-test but the use of a normative baseline that leads to a stronger test of prediction 1. This includes predictions about the pre-test but it also includes the *combined* effect of prediction 1 (prior experience) and predictions 2a,b (exposure), which had not been tested at the level we aim for.

Second, that learning depends on the amount of exposure. The authors are right, this is not typically tested in perceptual adaptation in speech (though there are at least a few where it is). But I don't see any models that really predict anything differently. Isn't this just a version of Thorndike's law of exercise or law of practice? Similarly, they argue that adaptation depends on the distribution being learned. OK, but isn't that the definition of distributional learning? And that learning is rapid. That's actually pretty interesting, but as the authors acknowledge (page 38-39) a lot of prior studies show that too.

We have completely restructured the introduction to be clearer about our contributions. The same holds for the Methods and Results sections, which now hopefully are much clearer. We agree with the reviewer that there is a lot of evidence that is consistent with distributional learning. And it seems, like the reviewer, we consider distributional learning the most elegant (though perhaps not most parsimonious) explanation a large number of findings about speech perception. But there are reasons why others disagree. As the revised introduction clarifies, not everyone believes that listeners adapt by learning phonetic distributions. And, to be fair, most existing works (including our own) provide rather indirect tests of that prediction. The revised introduction now makes this clearer. This also includes a discussion of what is known from what type of paradigm, following a suggestion the reviewer makes below.

Third, the authors point to the idea of diminishing returns - that learning will slow with more practice. However, this is also known as the power law of learning and has been shown in every domain of learning since the 1980s at least (Anderson, 1982; Logan, 1988; Newell & Rosenbloom, 1981) They point to this as a critical prediction of error minimization learning or prediction error, but this really falls out of a million forms of learning . For example, they write on line 905 "…this would raise questions as to whether similar predictions follow from other distributional learning accounts (e.g., C-CuRE normalization, McMurray & Jongman, 2011; exemplar models, Johnson, 1997; DNNs, Magnuson et al., 2020)…" I can't speak for C-Cure which assumes that distributional learning has happened, but doesn't posit a mechanism for that. But certainly exemplar models would show this - the classic work on the power law of learning was pitched in terms of instance- or exemplar-models (Logan, 1988; Palmeri, 1997) and both Palmeri and Logan offer a lovely mathematical treatment of how the power law is almost an unavoidable consequence of these architecture. Similarly, DNN's like Magnuson's are explicitly based on back-propagation of error, which in turn is based on minimization of prediction error (using essentially the delta rule or the Rescorla-wagner rule). So both of those frameworks would almost certainly show the same effects. The point is that I'm not sure that demonstrating that perceptual learning in speech also shows this effect is all that unexpected.

We agree. The revised introduction now clarifies that diminishing returns are predicted by many theories and have been found for many learning phenomena (citing many of the papers the reviewer kindly provided here and below). This does not, however, mean that we should take it for granted for rapid adaptation in speech perception. In fact, as we now clarify, one question about rapid adaptation is precisely whether it involves learning (as opposed to e.g., changes in decision-making).

Now, the one place that is super novel and super interesting is the fact that the learners do not appear to fully learn the new shifted categories - they seem to stop before they get all the way there. That's novel and hasn't really been shown. And frankly the sophistication of the ideal observer models trained on the same data as the subjects are makes it starkly clear that the subjects are doing this by providing a very clear view of what the subjects "should" be doing. That's probably the most compelling aspect of the paper. But to get there you have to wade through [sic]

We appreciate this clear feedback. We agree with the reviewer that the paper needed restructuring, and have done so (see overview at the top of this letter). In particular, we now state up front that prediction 4 about the convergence against the exposure statistics (formerly prediction 3) has not previously been tested. (For what it is worth, it’s basically impossible to test that prediction without also testing predictions 1 and 2a-b over an incremental paradigm, and this in turn means that one gets a test of the power law of learning—prediction 3—for ‘free’).

We hope it is ok to provide some additional perspective, as it is a common fate in our experience for more in-depth studies to elicit a lot of interest and engagement from reviewers, but then ultimately being delegated to ‘specialty journals’ for supposed lack of novelty.

Given the reviewer’s comments, we considered focusing the paper on the prediction R3 highlights. Like R3, we find this to be the most exciting result. We did, however, decide against this. Instead, the paper with all its presentational changes still embraces complexity. We think this is good and adequate. We believe that it is precisely this level of evaluation that is necessary to move the field to the next stage.

The reviewer might agree or disagree, but we hope they can see why some, like us, see the type of approach we present as an exciting new way to study adaptive speech perception and similar phenomena. Our goal is to show the field that there are important insights to be gained by carefully evaluating *predictive* models—i.e., models that make quantitative predictions that can be compared against gradient human behavior. It is only with such comparisons that we begin to see what the model does and does not get right, e.g., that distributional learning can account for 96% of the changes in listeners’ behavior and yet get some aspects of the data systematically wrong.

Major Concerns

\* To some extent it feels like the authors are sort of setting up the easy predictions to test. They do a nice job of laying out relevant principles like the effect of exposure amount, the role of prior expectations and so forth. But in exploring the permutations of these things they come to some fairly simplistic possibilities. For example, the prediction of diminishing returns is held up as a pretty important one that any model of learning needs to be able to show. That's kind of true. But this is a property of virtually all learning - in fact, its been termed a "law" (the power law of learning). Almost any kind of learning device will show it confronted with almost any learning problem. Similarly figure 1 proposes linear learning as a possibility? Really? People just increase incrementally and then suddenly stop when they hit the target? I don't think any model has ever posited that - there's always a slow down (back to diminishing returns?). And one -shot learning? That's been posited in areas like word object mapping where there's a ground truth in the world to refer to (unlike speech perception where there's a ground truth that is only known by the talker), but I don't see any plausible theories of speech perception that would posit this (particularly given the inherent noise in productions. I appreciate the rigor of the analysis here, but it risks setting up trivial predictions that do not really distinguish models.

We now clearly state that some of the predictions are more plausible than others. We seem to share similar theoretical biases/assumptions/takes on the literature with the reviewer. However, the fact is that not everyone takes for granted that rapid changes in speech perception involve *any* form of learning (see revised introduction). Many of the alternative theories are, admittedly, under-developed. But it is unclear whether rapid changes in e.g., decision-making could proceed linearly or instantaneously. Figure 1 is meant to illustrate these logical possibilities. Additionally, Figure 1 serves to onboard non-experts.

As for the reviewer’s point about the ground-truth, it is unclear to us how “a ground truth in the world” is different from a ground truth “only known by the talker”. In both cases, learners need to draw inferences. One might argue—and perhaps that’s what the reviewer has in mind—that listeners might have more uncertainty than word learners given the high degree of cross-talker variability. But this misses that listeners already have experienced many different talkers, and might be able to infer (rather than learn, in the more narrow sense of the word) which mixture of those previously learned talkers provides a good model for the current input. This is, of course, the very analysis we propose in the general discussion as *one* plausible theory for the findings we obtain. Critically, with regard to the reviewer’s point, this means that listeners might already have quite a bit more information than an infant learning new words, making ‘one-shot’ learning/inferences a more plausible candidate than it might at first blush appear to be to some readers.

\* I was surprised that no power analysis or justification was given - particularly given that part of the study was preregistered. I don't think every sample size needs to be justified by a priori power - particularly the first study in a new paradigm where effect sizes can't be known. I'm fine if the argument is just "we ran a lot because we didn't know what to expect". But even then it would be very useful to include a sensitivity/minimum detectable effect analysis to help the reader understand what kind of effects could be detected.

We appreciate this call out. We have not (yet) added a sensitivity analysis. But would be happy to do so if the reviewer believes this is critical.

To clarify our perspective on the value of power analyses for *this* study: standard power analyses would have been uninformative given that this is the first study of this type (for issues with the common practice of conducting power analyses over assumed effect sizes, such as the power to detect a “moderate” effect, see Xie et al., 2023). Instead, we had the benefit of model-guided predictions—we knew exactly how far listeners’ categorization functions should shift if they fully converged against the input (which they did not). And we had previously run many (non-incremental) distributional learning studies in our lab (with power analyses) and thus knew how many participants would give us reasonable power to detect even half of the shifts expected for an idealized learner (citations omitted for the sake of anonymity).

We also note that we find all tested hypotheses—all of them dictated by distributional learning models—confirmed (except for the prediction that learning should continue until convergence against the exposure statistics, for which we find decisive, strong rejection of the hypothesis). In short, our data is strikingly unambiguous despite the many (planned) questions we ask. This is not the type of patterning one would expect from an unpowered study.

What do power analyses add at this point? An informative analysis would require simulations over a wide range of effect sizes, and it would tell us that there are some effects that are small enough to give us little power in this paradigm. But what will other researchers make out of this? The minimally detectable effect size (e.g., the minimal change in the PSE for which we would have 95% ‘power’) should not be meaningful to a researcher who does not also know what shift to expect given a) the phonetic distributions in listeners’ prior experience, b) the phonetic distributions in the exposure input, c) the amount of exposure, d) the position of the test stimuli, and e) participants’ rate of attentional lapses (which presumably depends on the stimuli, task, and instructions the researcher will employ, etc.).

In short, we sympathize with the call for clear power estimates (and provide them in most of our work). However, such estimates carry little information—or even risk being misleading—in a field for which researchers have no good basis to calculate *expected* effect sizes (see also Corretta et al., 2023; Martin & Guest, 2021; Yarkoni & Westfall, 2017; Xie et al., 2023 for discussions of related points)! This is one of the many reasons why our revised introduction aims to emphasize the importance of quantitative models.

\* The other surprising omission. While there is a fairly strong theoretical motivation (despite my first concern), by the time I got to the methods, I really didn't know how the hypotheses mapped on to the experimental contrasts and/or conditions? Actually, in retrospect, I'm not entirely sure what the hypothesis were?! (which is odd considering how theoretical the intro was). It would help to have some clear statements of the form: "if perception works this way, then we should see a difference in [something] between [some two conditions]". By the time I got to the results, I just kind of had a vision of a general purpose, well constructed distributional learning task, and the authors were gonna just kind of see what it showed. But I don't think that's what they're up to here.

That is correct and we appreciate the charitable take. The revised methods section now more clearly links design decisions to our goals. These goals are stated in a new section in the introduction.

\* I never really understood how they are going to test the hypothesis that the prior state of the category system constrains or predicts subsequent adaptation. That seems to be one of their clear goals here. But they only kind of argue that because they had a pre-test they must be doing that . But the presumption seems to be here that everyone had the same prior state. And if everyone is the same to start, and everyone adapts the same, how does this test the hypothesis? To test this, it would seem like you'd need to manipulate the prior state and showed that this influenced the course of later learning? Maybe this could be done in an individual differences framework (e.g., compare the course of learning for people who's pretest boundary was a little lower to those with a slightly higher boundary).

We apologize that this was unclear. The revised introduction now points out explicitly that—unlike previous work—we test whether the phonetic distributions a ‘typical’ listener of US English would experience can predict listeners’ pre-test behavior. So, the reviewer is correct: it’s not the pre-test itself that is the innovation; rather, it is the use of the ideal observer and adaptor models that capture listeners’ prior expectations that is novel. The pre-test is a prerequisite in order to be able to test those predictions. This is now stated clearly in the Methods section.

(So, no, we are not yet modeling individual differences in experience; rather, we are adding a long overdue simple test as to whether distributional learning theories actually correctly capture listeners’ behavior at the start of the experiment *and* the way in which these expectations change with exposure. To anticipate a concern the reviewer might have: yes, if one—unlike recent reviews of the field—takes it for granted that distributional learning must underlie rapid changes in speech perception, this is not surprising. One might also argue that other work has found that at least the VOT distributions of US English qualitatively predict listeners’ behavior in non-adaptive perception experiments, e.g. Kronrod et al., 2016. We agree. We test whether those same prior expectations are compatible with the adaptive behavior listeners’ exhibit across blocks, once prior information is integrated with information from exposure. We find that simple distributional learning models that integrate prior and exposure information incrementally without additional constraints do *not* predict listeners’ behavior correctly, as they do not predict ‘premature convergence’.)

Minor Concerns

(For the most part these are not issues that drove my overall evaluation of the paper, but I bring them up as helpful suggestions

\* Line 29-32: The intro starts talk about how challenging speech perception is, but then ends with "Yet, listeners typically recognize speech quickly and accurately across a wide range of talkers and acoustic conditions…". This is a fairly standard way to introduce a cognitive science paper: explain how hard the problem and then present the mystery of how most people solve it. I've used it myself maybe 20 times or more! But I don't really buy it anymore. Something like 10-20% of people have developmental language disorder or dyslexia (both of which impact hearing loss). Speech perception doesn't fully develop until you are 20 (so that rules out most kids), and it starts declining in your 60s maybe (ruling out many adults), and then there is hearing loss. And bilinguals? Bilinguals perceive and adapt to speech differently in both the L1 and the L2. That's like 60% of the population. I know this whole paragraph is a kind of throwaway line to pique the readers' interest, but the fact is, that many of not most people probably don't solve this problem effortlessly. Its true that a slice of people can, but by framing it this way, it seems to artificially restrict the domain of cognitive science to just these perfect highly literate monolingual young adults.

Fair point. We agree with the reviewer that it’s good to avoid tropes. We have changed the sentence to “healthy young adult listeners …”. We’re open to additional suggestions (e.g., perhaps “neurotypical” would be preferred?).

\* Figure 1A, B: the long dashes make it really hard to read the figures? Maybe dots? Or dash-dots?

Thanks. We have followed the reviewer’s suggestion.

\* Figure 1D: Given that VOT is on the X axis of the top panels, would these work better transposed with VOT on the X axis? One less mental rotation for the reader.

We appreciate the reviewer’s suggestion. However, the remainder of the paper plots PSEs on the y-axis (anything else would force us to plot blocks along the y-axis, which would feel even less intuitive).

\* Line 94: "both error-driven theories (Harmon et al., 2019; Olejarczuk et al., 2018; Sohoglu & Davis, 2016) and theories of ideal information integration (Kleinschmidt, 2020b; Kleinschmidt & Jaeger, 2015) predict that adaptation initially proceeds quickly and then slows down as the listener approaches the correct mapping from the acoustic signal to phonetic categories (prediction 4 - diminishing returns)." This is not just a property of models of speech perception - "diminishing returns" is a central feature of the power law of learning (Anderson, 1982; Heathcote et al., 2000), and almost all associative theories of learning (e.g., from the animal learning literature) (Rescorla, 1988). This is not a huge problem, but given the ubiquity of diminishing returns in virtually all of learning, it begs the question as to whether this particular aspect of perceptual learning of speech really requires us to test it empirically, and if it needs an explanation,

Thank you for the reference. We have addressed this point above, and revised the introduction accordingly.

\* Line 314: "Each exposure block consisted of 24 /d/ and 24 /t/ trials,…" This didn't make sense to me at first - if it's a continua, how can you be sure what a /d/ or /t/ trial is. Later on it's clear to me that this is meant in a sort of mixture model way - first select which underlying phoneme it is, and then select (randomly) the observed VOT. But that's not clear yet. Might help to clarify that.

Thank you for pointing this out. The three conditions are meant to simulate three talkers with their different realizations of /d/ and /t/, determined by the placement of the phonetic distribution along the VOT continua. We have added additional clarifications to this section. We hope that the figures (also revised) are helpful, too.

\* Line 326: Why are these expressed as variance, not SD? I read them as SDs first (and I think most phonetically minded folks would read them this way) and as a result the estimates seemed huge (an 80 msec width in the /d/ distribution!) until you realize that they are squared (8.8 msec width is perfectly appropriate).

We now report SDs and variance (the latter are usually reported in distributional learning paradigms, e.g. Clayards et al., 2008; Kleinschmidt & Jaeger, 2016; Theodore & Monto, 2019).

\* Page 15, top paragraph: Distributional learning - at least as that term has been used in the literature - almost always refers to \*unsupervised\* learning. But suddenly we get these labeled trials, and it is clear that there is a supervisory signal too. I recognize that semantically, supervised distributional learning is quite sensible (you can use the supervisory signal to help learn the supervision). But this is not what the field is likely to expect from the term. It likely also affects learning in fairly dramatic ways (supervised and unsupervised learning are widely seen to have pretty different properties).

We now clarify in the introduction that we use both labeled and unlabeled input. The reviewer is correct that many researchers interested in distributional learning—in particular, those working on language *acquisition—*have focused on unsupervised learning. For acquisition, any reasonable distributional learning theory needs to involve unsupervised learning.

This contrasts with adaptive speech perception in adult, which requires remapping (through distributional learning, changes in decision-making, or other mechanism) the unfamiliar input to *familiar* categories. As we discuss in the *Methods* section, this adaptation might involve both supervised and unsupervised learning—given that the context often, but not always, effectively labels categories in everyday speech perception.

While we acknowledge the concern that supervised and unsupervised learning mechanisms may have distinct properties, existing evidence does not consistently show that the presence of labels fundamentally alters learning in the type of paradigm we employ—presumably because unlabeled exposure to bimodal distributions contains quite a bit of information for listeners who are already familiar with the relevant phonetic contrasts (e.g. Goudbeek, 2008; Kleinschmidt et al., 2015; Chladkova et al., 2017).

One of the things I didn't like about the introduction is that all perceptual adaptation paradigms are kind of treated the same: lexically guided retuning is the same as distributional learning. But they're not. I mean maybe if you're a fully committed Bayesian, they're all just means to get to the underlying statistics. But that's a strong assumption. And then when we get to this new semi-supervised paradigm here, the importance of the differences among learning paradigms really moves to the forefront. I think the manuscript would be much stronger if a) the introduction actually discussed the different paradigms; and b) this particular hybrid was foreshadowed earlier than in the methods.

We agree with the reviewer that it cannot be taken for granted that all paradigms tap into the same mechanisms. The revised introduction now makes this much clearer. Indeed, the goal to test all four predictions in a single paradigm is in part motivated by the fact that existing evidence is scattered across *different* paradigms (as we now clarify). The reviewer is also correct that considerations about the pros and cons of different paradigms informed our own choices, and this too should now be clearer. As in the previous submission, we continue to discuss the different paradigms in the general discussion (which continues to clearly distinguish between them).

\* Line 344: I really had a hard time following the design here. Were all participant exposed to both the +10 shifted and +40 shifted blocks or was that between subject? What is this block order factor? Did everyone get baseline then one of the shifts? Or did people just get a single thing (baseline, +10 or +40)? I think part of the issue is that some of the randomization stuff isn't really that important (e.g., since Gorilla can't randomize on the fly, there were multiple lists for different subjects in a condition) and others were really important and the design section doesn't really distinguish them. But also the step through of the various "phases" is embedded in the procedures and you really have to work at it to understand the bigger structure - it might also be useful to have some kind of simple statement first (e.g., people got a pre-test, a training, then a post-test) or a visualization of the flow.

We apologize that the experiment design was difficult to follow. In addition to the caption of the design figure, we now also state in the text that exposure distributions were manipulated *between* participants. Each participant only saw one exposure condition. (Perhaps some of the confusion was due to the naming of conditions, which now has changed. There is nothing special about what we previously called the “baseline” condition.)

We also revised the *Procedure* section to clarify that “block order” refers to the Latin square design that carefully counter-balanced the VOT samples within and across conditions in a way that combines two innovations: (1) as described under Procedure, we *sampled* stimuli during exposure, avoiding the highly unnatural, perfectly symmetric exposure distributions that have been used in previous distributional learning experiments (experiments that explicitly manipulate the exposure *distribution* in known and controlled ways), but we also (2) hold this sampling constant across exposure conditions, removing any potential confounds it could introduce. This way, the three exposure distributions are exact copies of each other, except that they are shifted by a constant amount.

As we now clarify, this counterbalancing is also shown in the design overview figure (Figure 2), which continues to anticipate the entire design and procedure. This figure summarizes the entire test-exposure-test-exposure-test-exposure-test-test-test structure. We have updated it somewhat to clarify the design. Finally, the same careful counter-balancing is shown in Figure 4, and is described in its revised caption. If the reviewer think that it would be better to move some of these details into the SI, we would be happy to do so.

Finally, we note that we did not use Gorilla, and that we had no limitations with randomization. Rather, the randomization we describe reflects the attention to detail that we believe to be important in achieving increased ecologically validity while maintaining the level of control that is necessary for a fair quantitative evaluation of actual models. A welcome (intended) side effect of all of this effort was that the exposure statistics differed not only between exposure condition (between participants) but also differed somewhat from block to block *within* exposure condition (see Figure 4). This gave us more unique combinations of different exposure statistics against which we could fit the ideal adaptor.

\* Page 17. I really love the use of a model which embraces lapse rates. However, I two minor concerns and a question. Lapse rates capture differences at asymptote, and the authors are right that if you don't capture them you risk getting the boundary wrong. One minor concern is that a lot of readers won't be familiar with the function or the standard parameter names - it would probably be a good idea to define the lapse concept more clearly. But here's the second concern, lapse rate may not be the right term (even though that's the standard term for that variable). The term derives from detection paradigms where you might "miss" a stimulus due to a "lapse" of attention. This miss is expected to be independent of the x axis which is why it affects asymptote. However, in a categorization paradigm, a difference at asymptote might not be a lapse of attention - it may be that people are overall biased to report one category, affecting the asymptote at one category but not that the other), or that nothing sounds like a good /d/ or /t/ to them (affecting both asymptotes equally). It might be helpful to adopt a more neutral term to describe asymptotic differences (even as one wants to acknowledge that the traditional term is lapse). Finally, some psychometric functions would put the lapse rate on only one side of the transition (e.g., for a detection paradigm, people are assumed that they'll always detect the loudest stimuli, but lapses will affect the asymptote for the quiet end). That's obviously It would be helpful to be clear that you used the four-parameter function with lapses on both ends (assuming you did).

We appreciate these points. Our psychometric mixed-effects model continues to be described in detail in the SI (as indicated at the top of the results section). Given that, as the reviewer also pointed out, readers already have to “wade” through many details, we have decided to keep those details in the SI.

We carefully formulated our mixed-effects psychometric model to avoid the issues anticipated by the reviewer. In particular, we are using a mixture model formulation, not the standard non-linear model formulation that readers of e.g., Wichman & Hill or Prins would be familiar with:

lapse \* response\_bias + (1-lapse) \* perceptual\_model

This mixture formulation is computationally more stable (better convergence) and also more suitable for our purpose in terms of its parameterization. E.g., our lapse rate is indeed a lapse rate. The bias is modeled separately (well, for the main analyses, we set it to uniform, in line with what every other analysis in the field implicitly assumes; but we confirmed in additional analyses that this assumption was justified).

After talking with our vision colleagues, we think that our paper might well be one of the first papers that presents a single mixed-effects psychometric analysis across blocks and conditions (e.g., psignifit and similar matlab packages used in vision research do not offer the functionality we present). If the reviewer thinks this deserves highlighting, we can do so.

\* The Bayesian rather than frequentist analyses are quite appropriate, but still not widespread in the field. It might be helpful to remind the reader the typical ranges of bayes factor and how to interpret them.

Great suggestion. We have added this information at the start of the results section.

\* The idealized learner model is really helpful in understanding what the subjects are doing, but it's a bit hard to track because the results are pitched verbally in terms of PSE change and visualized in terms of "accuracy" (which is a less useful construct in this kind of categorization where there is no ground truth).

The idealized learner is shown both in the PSE figure (Figure 6c) *and* in the “accuracy” figure. This is stated both in the text and in the captions of the respective figures. We have revised the figures to make this even clearer.

And just to reiterate our point from above: there *is* a ground truth (each token is drawn from either the /d/ or the /t/ distribution, just like in natural speech. That *is* the ground truth).

It seems to me that a visualization more like Figure 6C might be more valuable - to see the boundaries from the listeners alongside the ideal boundary from the model as they unfold over time…. Oh wait…something like this is provided in Figure 8, ten pages later. Might be helpful to make a version of this earlier - it would be cool to compare the versions with the ideal adapter vs. the ideal observer.

Following the reviewer’s advice, we have moved Figure 8 (and the surrounding text) to much earlier in the paper. We note that it cannot be meaningfully compared against the ideal observer model. Only the ideal adaptor is a learning model, whereas the ideal observer models simply describe what an observer should do that has full access to the exposure distributions. This is now clearly stated both at the start of the Method section, and then again when we first introduce the ideal adaptor.

\* The authors criticize the use of synthetic speech in multiple places, but I'm not sure its really all that widespread. My recollection is that most of the lexically guided retuning work with fricatives, for example, uses natural recordings and techniques like sample averaging to create the stimuli which sound highly natural, and most of the more recent VOT studies do a type of cross-splicing similar to what is done here. In fact, I'm not even sure if any of the existing Klatt synthesizers even work in the latest versions of windows. One can critique all of these stimulus construction techniques on phonetic grounds (sample averaged, in particular, comes with serious issues for this purpose), but I'n not sure sounding robotic is one of them. I could be wrong - -I've read most of these papers, and didn't really keep a catalog of stimulus construction types - but I'd be careful with this assertion.

The reviewer is correct that there are more recent distributional learning studies that have used less robotic sounding stimuli. We were referring to specifically distributional learning studies on VOT, of which only Theodore and Monto used more naturalistic stimuli. But we have removed this sentence from the Method section.

\* The authors use the term PSE (point of subjective equality) as the key DV in a lot of analyses. That seems like the right one, but why use that term? It's a fine term, and it comes out of the psychometrics literature, but everyone in speech would call it the category boundary? Why make your audience learn a new term? I don't see the point and it's a bit off putting (or even haughty).

The caption of Figure 1 now states that the PSE is “aka category boundary”. Different readers of Cognition will expect different terminology. We were trying to avoid the use of category boundary because it can be a problematic term. To some researcher, the term “category boundary” refers to a boundary in the phonetic space determined solely by the cue distributions of each category. The PSE, however, is a function of both those distribution (the likelihood in Bayesian terms) and the context (the prior in Bayesian terms).

Specifically, we imagine that the common use of the term category boundary is part of the reason why some researchers seem to consider terms like “boundary shift” to be an *explanation* rather than a *description* of adaptation results (to us, “PSE shift” would seem to avoid this misinterpretation, as it unambiguously refers to the categorization behavior). We are trying to avoid this type of language (which perhaps *is* haughty …), and we are also trying to make our findings accessible to non-speech researchers (which hopefully isn’t haughty). If the reviewer thinks our concerns are unfounded and the use of the term PSE makes things unnecessarily difficult for speech perception researchers, we will remove it from the paper (neither of the other two reviewers took issue with it).

\* I really like the ideal adaptor models but I found it very hard to understand how it is different than the various other ideal Bayesian models that were presented earlier. It would help the reader to have a more explicit compare and contrast - maybe even a diagram.

We now introduce this model in more detail in an earlier section. There, we also clarify how it differs from the ideal observers (which are *not* learning models).

\* Section 4.3.3 seems to come too late. It might be more effective to present that before the ideal adaptor model. That is, put the concerns with premature convergence to rest before you present an explanatory model.

This whole section has changed because we moved the ideal adaptor to a much earlier point in the paper.

\* Line 1065: "While it is difficult to evaluate this explanation without a specific model of how listener learn from unlabeled tokens, one consideration suggests that it is not sufficient to explain our data…." McMurray, Aslin, et al. (2009) have a mixture of Gaussian's model that does distributional learning from unlabeled exemplars…. That could be a promising avenue for future exploration.

Correct. Thanks for spotting this. We now cite McMurray et al further down in the paragraph. Together with Vallabha et al., it’s indeed one of the earlier unsupervised models.

\* Line 1065. The other thing that struck me about this statement though, is that the authors appear to be dramatically minimizing the role of unsupervised learning. That comes out now??? Distributional learning was originally posited by Jessica Maye to be an entirely unsupervised process that infants may use to acquire the early phoneme categories of their language (Maye & Gerken, 2000; Maye et al., 2003). In the history of that approach to learning, the unsupervised nature is the core. There have been tons of computational models of this from connectionist (Gauthier et al., 2007; Guenther & Gjaja, 1996; McMurray, Horst, et al., 2009) and non-connectionist (McMurray, Aslin, et al., 2009; Toscano & McMurray, 2010) approaches. And clear demonstrations that humans can do both phonetic category adaptation (Clayards et al., 2008) and learn new categories (Escudero et al., 2011; Escudero & Williams, 2014; Goudbeek et al., 2008; Goudbeek et al., 2009) without feedback. It feels oddly revisionist to claim here - particularly this late in the paper - -that the supervised portion is what's driving the show.

We appreciate the reviewer's perspective but believe there may be a misunderstanding regarding our intended argument. We do not claim that the supervised portion is the primary driver of learning as that is not what we believe. However, we recognize that our original phrasing may have led to this interpretation, so we have revised the sentence to read: "While it is difficult to evaluate this explanation without a specific model of how listeners learn from unlabeled tokens with an uninformative (uniform) distribution over the phonetic cues…" to clarify our position.

Beyond the many references the reviewer provides, there are also several papers that directly compare unlabeled to labeled or semi-labeled exposure. We now cite these works in the Methods section. They do not find any difference between the two conditions (at least for the much longer exposure they use, compared to ours).

Critically, all of these works are about *informative* unlabeled exposure, i.e., exposure to multimodal (typically bimodal) distributions. To our knowledge, prior work has not modeled learning from uninformative unlabeled exposure, such as the uniform distribution of test tokens discussed in the paragraph the reviewer references. Our statement therefore was not intended to be revisionist but rather a clarification of the specific learning conditions under consideration.

\* The authors do a very nice job of evaluating their own statistical models to ensure that the priors aren't creating an effect that isn't there (in particular the premature stopping). But given all this, I wonder if they should consider (as a secondary analysis) a non-Bayesian approach. It seems like that's a big part of the problem. But if they moved to a two parameter logistic (which they admit is probably fine, since the lapse rates were minimal) they could do it in a standard mixed model? Or maybe avoid mixed models all together (there aren't any random items here) and do some kind of curvefitting approach? I don't think either of these are superior to what they are currently doing, but it could offer reassurance that the priors in the current psychometric approach aren't driving the effect.

One of the Bayesian auxiliary analyses we conduct employed a uniform prior, removing any bias from the estimation of parameters. As we now clarify, this is essentially identical to a frequentist analysis. That analysis replicated all findings we report (see SI 5.1). In short, there is no problem here.

Given this context, we hope it is ok to say that we see little value in adding frequentist analyses to the paper. First, frequentist models would likely not at all converge with the full random effect structure (the ‘dark secret’ of those models that has prompted dozens of highly cited papers in the psych sciences). Second, while there are some libraries for frequentist psychometric models, they all have limitations with regard to the designs they allow, whereas the bmrs library has no such limitations (and switching to ordinary logistic regression seems like a step back, risking that readers will miss that this would just be ok because we found very low lapse rates). Third, we employ hypothesis tests that would be hard to transfer into a frequentist model without refitting the model in many different ways. Finally, we note that there are random effects, both for subject and for items.

We thank the reviewer for the careful review, and the particularly constructive criticism. We also appreciated the list of references, most of which we have integrated into the text.