For general letter:

* Need to move some of the novelty explanations into the main part.
  + There is lots of evidence that is qualitatively compatible with the theory of distributional learning. But most tests are weak.
  + It’s not that we think the idea to test incrementally is novel. What is novel is to begin to assess the effects of *distributional exposure* incrementally during the earliest moments of exposure. And this question is important and remains unaddressed.
  + I think this is best achieved by removing the claim that the paradigm is novel, but we still need to get across that the idea is novel.
  + Drive home that our paradigm is much more fine grained --- both in terms of unique stimuli during exposure and test, and the method used to estimate changes in cat function.
* Need to get across that there is a huge literature and that, while we now anticipate some of the themes more clearly, we found it feasible to only discuss the themes in more detail in the general discussion.
* Let’s discuss whether we want to change the condition labels. In a way reviewer 2 gave us license to do that. We can still be clear that the labeling was arbitrary prior to the IO development.
  + We need to make clear that we developed the Ios in parallel while preparing the experiment. It took us about 1 year to implement and test all parts of the Ios (reported in more detail in the SI). Nobody had really deployed such IOs, and evaluated them through norming studies. (use this to clarify the novelty of this aspect).
  + Ideally, we would have done that first and then carefully selected our conditions based on that, but we had to balance the time pressure inherent in a PhD thesis against those considerations.
  + We thus tried our best to have conditions that center around what we expected to be the baseline based on norming studies.
  + The use of the label “baseline” was unfortunate. We should have simply labeled them A, B, C, as we did not know exactly how exactly these conditions would be shifted relative to listeners’ priors. That only became clear once we had developed the IO or once we saw how they shifted relative to pre-exposure.

Best, C. T. (1995). A direct realist view of cross-language speech perception. In. W. Strange, Speech perception and linguistic experience: Theoretical and methodological issues in cross-language speech research (pp.167-200).

Escudero, P., Benders, T., & Wanrooij, K. (2011). Enhanced bimodal distributions facilitate the learning of second language vowels. Journal of the Acoustical Society of America, 130(4), EL206-EL212.

Kraljic, T., & Samuel, A. G. (2007). Perceptual adjustments to multiple speakers. Journal of Memory and Language, 56, 1-15.

Poellmann, K., McQueen, J. M., & Mitterer, H. (2011). The time course of perceptual learning. In W.-S. Lee, & E. Zee (Eds.), Proceedings of the 17th International Congress of Phonetic Sciences 2011 [ICPhS XVII] (pp. 1618-1621). Hong Kong: Department of Chinese, Translation and Linguistics, City University of Hong Kong.

Witteman, M. J., Weber, A., & McQueen, J. M. (2013). Foreign accent strength and listener familiarity with an accent co-determine speed of perceptual adaptation. Attention, Perception & Psychophysics, 75, 537-556.

Zhang, X., & Holt, L. L. (2018). Simultaneous tracking of coevolving distributional regularities in speech. Journal of Experimental Psychology: Human Perception and Performance, 44, 1760-1779.

Bertelson, P., Vroomen, J., & de Gelder, B. (2003). Visual recalibration of auditory speech identification: A McGurk aftereffect. Psychological Science 14(6): 592-597.

Logan, J. S., Lively, S. E., &Pisoni, D. B. (1991). Training Japanese listeners to identify English /r/ and /l/: A first report. Journal of the Acoustical Society of America, 89, 874-886.

Eisner, F., & McQueen, J. M. (2006). Perceptual learning in speech: Stability over time. Journal of the Acoustical Society of America, 119(4), 1950-1953.

As the revised manuscript clarifies there are many findings that are *compatible* with the hypothesis that distributional learning explains the observed effects. However, as has been worked out in detail in recent reviews of the field, there is a considerable difference between such qualitative compatibilities, on the one hand, and paradigms that lend themselves to the quantitative evaluation of distributional learning hypotheses (see Xie et al., 2023; also Guest & Martin, XXXX; Yarkoni & Westfall, 2017).

To just name one example, Xie et al (2023) show that qualitatively completely different theories can account for several of the signature results sometimes associated with distributional learning. A key goal for the present work was thus to move towards a paradigm that lends itself better to careful *quantitative* evaluation of distributional learning theories by manipulating—under detailed control of the experimenter—both the amount and type of exposure. This is *precisely* what recent reviews asked for, and we show how this type of paradigm allows us to evaluate models of learning at a level that they have *not* previously been evaluated (see the ideal adaptor model in the general discussion).

For instance, under a standard way of analyzing our data (incrementally or not), all of our results would have appeared to follow the predictions of distributional learning theories. It is only because of the combination of a) experimental control over the incrementally presented distributional exposure and b) the use of computational analyses (ideal observers and adaptors) that we were able to identify that the results do *not*, in fact, follow the type of distributional learning theories that have previously been implemented.

Finally, we note that Kraljic & Samuel (2007) do *not* manipulate the amount of exposure independent of talker identity. Rather, they address another (very interesting) question about whether listeners adapt separately to different talkers. There is a whole separate line of research on this question (more work by Kraljic & Samuel, but also Saltzman & Myers, Kurumada and collaborators, etc.). Given that this question is somewhat independent of the more basic question we are addressing, we have decided not to also summarize that entire literature.

Point-by-point reply to reviewers’ feedback. Our responses are shown 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 timecourse/nature of adaptation for novel vowel shifts?

This is a great point. 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 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.

1. 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.** The short of it is that **we have completely revised the introduction** to be clearer about our goals, and about the fact that it is *the combination of paradigm, analyses, and model-guided interpretation* that make the present study so different from previous work.

1. 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 we previously only raised in the discussion.** In this context, we also 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.

1. Some of the in-text citations have author initials (e.g. AA or M.M.)
2. 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.
3. Figure 6: the labels are overlapping in the 3 lower right corner panels - please fix

We fixed all of these points. Thank you for catching these mistakes!

**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 letter to the editor, 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.

Our design , combined with the use of analysis methods that provide quantitative estimates of p(t response | VOT step, exposure condition, exposure block order, current test block), allowed us to obtain **384 different measurement within and across participants for which the ideal adaptor model we test makes different quantitative predictions** (Test 1-4 x 3 exposure conditions with different underling distributions x 3 Latin-square designed orders of the specific random samples drawn from those distributions x 12 VOT steps during test). As we now clarify, assessing how much of the changes in participants’ perception at these 384 measurement points can be captured by a model of distributional learning (with only 3 DFs!) is a substantial contribution to previous work.

To appreciate just how different what we do is to previous work, it helps to consider that >99.99% of previous work has not tested *models* of distributional learning at all. While those studies have been critical in establishing qualitative compatibility with the distributional learning hypothesis, **previous work leaves completely open whether distributional learning can actually explain a non-trivial share of the observed changes in listeners’ behavior—as would be expected if distributional learning is the core mechanism underlying rapid adaptation during speech perception**.

Even the very few studies that have compared listeners’ behavior to quantitative models of distributional learning have been limited to qualitative comparisons of typically two exposure conditions after much longer exposure than in our experiment. We realize that the importance of quantitative tests remains under-appreciated in much of the field (in our experience, there is a tendency to dismiss them as ‘number crunching’). That is why **we now clarify just how much stronger—more able to identify limitations of existing theories—these quantitative tests are compared to previous work.** In our responses below, we clarify these points further.

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 letter to the editor, **we have revised the paper to be clearer that the naming of the conditions is 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 clearly specified, competing 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 is certainly plausible, it’s perhaps based on intuition or experience with previous experiments, rather than an existing model?

1. 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, and hope that it will make 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 predictions of distributional learning models.

1. 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.

**We have now integrated some of these papers into the introduction where appropriate.** We note though that all 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. 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. As we have addressed these points in the letter to the editor, we do not repeat them here.

1. 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. As our response above might already clarify, it is the combination of the paradigm, the psychometric model, and the learning models we employ that allows us to make the novel contributions we present. **The revised manuscript does not present the paradigm itself as novel.**

We hope the reviewer does not mind if we provide a bit more context here, as to what motivated our study. 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 different question from how *rapid* adaptation over mere minutes of exposure affects L1 speech perception (and in the absence of any reference to “training”: unlike in L2 learning studies, our participants were not asked to learn a new language; they were simply listening to someone speaking in the listener’s L1). While it is quite possible that rapid changes in L1 perception originate in the same mechanisms as L2 acquisition, and that this includes distributional learning, that is by no means to be taken for granted (see also 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 brain imaging research on adaptive speech perception is that rapid adaptation is achieved by changes in *decision-making* (REFS), 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 the work of Bertelson et al or Vroomen et al., 2007 (as well as Kleinschmidt & Jaeger, 2011, 2012; etc.). Like the work by Pisoni et al, these studies very much inspired the present work, which is why we cited many of them in the introduction, and returned to them in detail in the general discussion (we still do). But a*ll* of these studies study 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 paradigms draws on the same mechanisms that underlie adaptation to distributional exposure, and that this involves distributional learning (as hypothesized in e.g., Kleinschmidt & Jaeger, 2011, 2012). 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 come up if the stimulus presented on each trial is identical. Additionally, exposure in the studies cited by the reviewer is 100% labeled, and highly artificial sounding (one of us has worked with these stimuli in Kleinschmidt & Jaeger, 2011). We won’t repeat here the criticisms that have been leveled against such paradigms as a window into everyday speech perception (they are reviewed in detail in recent reviews by e.g., M. Baese-Berk). Neither do we mean to claim that our paradigm removes *all* of these concerns. We do, however, submit that there is a substantial difference between the approach employed in the present study and early work on visually- or lexically-guided perceptual retuning (as we continue to discuss in detail in our general discussion).

The importance to not taking for granted that different paradigms employ the same mechanisms is highlighted in several of the recent reviews we mentioned above (see also Cummings & Theodore, 2023 on lexically-guided perceptual recalibration, or reviews by Samuel between 2010 and now, which repeatedly emphasize that point).

Additional comments

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

We have added a single-sentence summary of the result. We also anticipate the result more clearly at the end of the introduction. We hope that is a fair compromise.

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. S**tandard 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. stimulus along the *dill-till continuum* with response options "dill" and "din". Anything else would have introduced another task-dimension into the paradigm, as participants would have had to decide what it means that 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, the stimulus was 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 other results from our lab strongly suggest that fully labeled or fully unlabeled exposure would not have qualitatively changed the results *in this type of paradigm*; see also Kleinschmidt, 2020). We have revised this paragraph somewhat 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 noting how this may confuse readers. We have added further explanation within the caption that for null tests BFs of > 1 indicate increased support for the null. We have also included the *Probability of Direction* (pd) as an additional index for assessing the evidence for an effect. The pd indicates the certainty of the estimated direction of an effect. The pd is calculated as the proportion of the posterior that falls within the sign of the measure of central tendency. In a null hypothesis test one would expect the pd to be roughly 50% -- half of the posterior falling within either side of 0, indicating support for the null.

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 so much for noticing these issues!

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 have now describe it before the general discussion. We also elaborate on this figure.** Yes, the CIs overlap with the model predictions (which is directly related to the fact that the ideal adaptor achieves such a high R2 of 96%).

However, there is a subtle, qualitative mismatch between the model’s predictions and listeners’ behavior: the model will always predict convergence with sufficient exposure (we confirmed this in separate simulations but it’s also follows mathematically from the way the model works), whereas listeners seem to plateau. Even great models can be partially wrong. **This shows in Figure 8, 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). However, instead of appealing to the Figure, **we now present additional Bayesian hypothesis tests that quantify this pattern.** These tests support our point.

**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. MORE HERE.

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.

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.

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. 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

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.

\* 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.

\* 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.

\* 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).

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.

\* 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). For what it is worth, this is also a common way of plotting data in other papers that have measured incremental changes in categorization (e.g., Vroomen et al., 2007; Kleinschmidt & Jaeger, 2015; Kleinschmidt, 2020; Cummings & Theodore, 2023).

\* 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 address 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 realisations of /d/ and /t/, determined by the placement of the phonetic distribution along the VOT continua.

\* 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 understand R3’s point to stick to terms and scales familiar to the target audience however in order to maintain congruence with previous studies of this distributional learning paradigm (e.g. Clayards et al., 2008; K&J2016; Theodore & Monto, 2019) we would prefer to keep the description in terms of variance. As a compromise, we have included the SD values when specifying the distribution in lines xxx and have edited the labels in Figure 4 to reflect the SD instead.

\* 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 distrubutional 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). 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.

\* 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)?

We agree that the experiment design may be difficult to follow without a close read given the between and within participants manipulations, and presumably the condition names. We tried to communicate as clearly as possible with Figure 2 through colour-coding and clear captions as well as when we refer to it in lines 124 -127. With that same objective in mind we had aimed to provide more detailed information about the stimuli between and within each condition through the histograms in figure 4.

After considering feedback from R2 and R3, we have added/edited the following:

1. The condition names now reflect the predicted PSE of each condition
2. The caption for Figure 2 now reads: “The three between-groups exposure conditions (rows) differed in … “
3. Reference to Figure 2 in l.xxx now reads: Between groups of participants, we manipulate the *distance between* the distributions of phonetic cues in the exposure input. The number of tokens that make up entire distributions within each group were evenly distributed between the three exposure blocks (48 tokens per block). This set up should be viewed as the exposure distribution being fully revealed by the end of exposure block 3 (see Figure 4 for more details)

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.

\* 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).

\* 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.

\* 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). 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.

\* 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 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).

\* 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.

\* 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.

\* 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.

\* 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.

\* 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. That analysis replicated all findings we report (see SI XXX).** In short, there is no problem here. We were just aiming to be very cautious.

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, many of which we have integrated into the text.

References

\* Anderson, J. R. (1982). Acquisition of cognitive skill. Psychological Review, 89, 396-406.

\* Clayards, M., Tanenhaus, M. K., Aslin, R. N., & Jacobs, R. A. (2008). Perception of speech reflects optimal use of probabilistic speech cues. Cognition, 108(3), 804-809.

\* Escudero, P., Benders, T., & Wanrooij, K. (2011). Enhanced bimodal distributions facilitate the learning of second language vowels. The Journal of the Acoustical Society of America, 130(4), EL206-EL212. https://doi.org/10.1121/1.3629144

\* Escudero, P., & Williams, D. (2014). Distributional learning has immediate and long-lasting effects. Cognition, 133(2), 408-413. https://doi.org/https://doi.org/10.1016/j.cognition.2014.07.002

\* Gauthier, B., Shi, R., & Xu, Y. (2007). Learning phonetic categories by tracking movements. Cognition, 103(1), 80-106. https://doi.org/https://doi.org/10.1016/j.cognition.2006.03.002

\* Goudbeek, M., Cutler, A., & Smits, R. (2008). Supervised and unsupervised learning of multidimensionally varying non-native speech categories. Speech Communication, 50(2), 109-125.

\* Goudbeek, M., Swingley, D., & Smits, R. (2009). Supervised and unsupervised learning of multidimensional acoustic categories. Journal of Experimental Psychology: Human Perception and Performance, 35(6), 1913-1933. https://doi.org/10.1037/a0015781

\* Guenther, F. H., & Gjaja, M. (1996). The perceptual magnet effect as an emergent property of neural map formation. Journal of the Acoustical Society of America, 100, 1111-1112.

\* Heathcote, A., Brown, S., & Mewhort, D. J. K. (2000). The power law repealed: The case for an exponential law of practice. Psychonomic Bulletin & Review, 7, 185-207.

\* Logan, G. D. (1988). Toward an instance theory of automatization. Psychological Review, 95, 492-527.

\* Maye, J., & Gerken, L. (2000). Learning phonemes without minimal pairs. Proceedings of the 24th annual Boston university conference on language development,

\* Maye, J., Werker, J. F., & Gerken, L. (2003). Infant sensitivity to distributional information can affect phonetic discrimination. Cognition, 82, 101-111.

\* McMurray, B., Aslin, R. N., & Toscano, J. C. (2009). Statistical learning of phonetic categories: Insights from a computational approach. Developmental Science, 12(3), 369-379.

\* McMurray, B., Horst, J. S., Toscano, J. C., & Samuelson, L. (2009). Towards an integration of connectionist learning and dynamical systems processing: case studies in speech and lexical development. In J. Spencer, M. Thomas, & J. L. McClelland (Eds.), Toward a Unified Theory of Development: Connectionism and Dynamic Systems Theory Re-Considered. (pp. 218-252). Oxford University Press.

\* Newell, A., & Rosenbloom, P. S. (1981). Mechanisms of skill acquisition and law of practice. In J. R. Anderson (Ed.), Cognitive skills and their acquisition (pp. 1-55). Erlbaum.

\* Palmeri, T. J. (1997). Exemplar similarity and the development of automaticity. Journal of Experimental Psychology: Learning, Memory & Cognition, 23, 324-354.

\* Rescorla, R. A. (1988). Pavlovian conditioning: It's not what you think it is. American Psychologist, 43(3), 151-160.

\* Toscano, J. C., & McMurray, B. (2010). Cue integration with categories: Weighting acoustic cues in speech using unsupervised learning and distributional statistics. Cognitive Science, 34(3), 436-464.