WHAT CHANGES WHEN WE TUNE INTO TALKER-SPECIFIC PROSODY?

ANDRÉS BUXÓ-LUGO
CHIGUSA KURUMADA
Department of Psychology, University at Buffalo
Department of Brain and Cognitive Sciences, University of Rochester

Abstract

A major puzzle of spoken language comprehension is how listeners navigate variability in speech prosody to infer the intended meaning, for instance pitch movements signaling questions vs. statements. Listeners must adapt to variability across talkers (e.g., according to their age, gender, accents) as well as linguistic contexts. The relationship between the phonetic variations and linguistic (e.g., syntactic) contexts that host the variations is not yet well-understood, however. Using resynthesized 11-step continua between rising and falling contours, we investigated the extent to which prosodic adaptation is conditioned on a syntactic context. Over two blocks, participants repeatedly categorized "It's X-ing" items (e.g., "It's cooking") sampled from the continua as questions or statements. In between, one group of listeners heard ambiguous items midway along the continua with different, interrogative syntax ("Is it cooking") and items near the falling end of the continua with declarative syntax ("It's cooking"). Another group heard the reverse. Prosodic adaptation was observed only in the listeners who heard the ambiguous prosody with the declarative syntax, congruent with the test tokens. We argue that the effective contextual conditioning facilitates robust prosodic adaptation while offsetting the risk of over-generalization across tokens with distinct phonetic features.

1 Introduction

Human language allows us to communicate highly complex and abstract information such as intentions and emotions. Speech prosody, defined as a holistic impression of melodic and rhythmic aspects of speech, plays a critical role in this process (e.g., Bolinger, 1964; Pierrehumbert & Hirschberg, 1990; Ladd, 2008; Cole, 2015; Cutler, 2015; Dahan, 2015). Prominence placed on a particular word can convey the talker's intention to signal a contextually salient contrast (e.g., "Olivia likes LEMONS (not oranges)" vs. "OLIVIA (not Annie) likes lemons"). Similarly, a rising terminal pitch of an utterance can distinguish a question from a statement. Complicating the listeners' task, however, realizations of prosody can vary substantially across talkers due to factors both systematic (e.g., vocal tract length, speaking rate) and incidental (e.g., speech errors) (Arvaniti, 2011; D'Imperio, Grice & Cangemi, 2016; Breen et al, 2018). Understanding the meaning of prosody must therefore include processes that aid listeners in navigating this talker-variability.

Recent studies have begun to uncover sources of flexibility necessary for robust prosodic processing. Neuroimaging work has suggested that the human auditory cortex encodes talker-normalized *relative* pitch in addition to absolute pitch (Tang et al., 2017). Furthermore, listeners' interpretations of coherent pitch movements (i.e., intonation) can further adapt to the statistics of the acoustic / phonetic features of a given talker. For instance, Xie, Buxó-Lugo, & Kurumada (2021) exposed three groups of listeners to distinct statistics of pitch movements signaling questions or statements. After only eight minutes of exposure, these groups of listeners derived diverging interpretations for novel, ambiguous tokens. Similar cases of rapid prosodic adaptation have been demonstrated for pragmatic (Patel et al., 2011; Ito et al, 2017; Kurumada et al., 2017; Nakamura et al., 2019) and affective (Woodard et al., 2021) meanings. What is as of yet mostly unknown is: *What changes in response to an accumulating exposure to a particular talker?*

The received wisdom in the literature suggests that prosodic processing calibrates to, or compensates for, a base vocal pitch and its range for a given talker. Beckman (1995) puts it as "all of our theories of intonational structure include at least an implicit representation of the speaker's overall pitch range in our models of the hearer's competence." Listeners may apply distinct expectations for vocal pitch in, say, female vs. male talkers (Gussenhoven & Rietveld, 1998; Bishop & Keating, 2003) or children vs. adults (Patel & Brayton, 2009). And such adjustment can happen rapidly (Lee, 2009). In real time as an utterance unfolds, an earlier part of an utterance serves as an anchor to shift expectations for subsequent parts (e.g., Dilley & Pitt, 2010; Saindon et al., 2017). As a result, listeners can flexibly map specific stimulus values to linguistically meaningful prosodic categories (e.g., rising vs. falling pitch) at different rates across talkers.

What has not been directly tested is whether listeners adapt more specific, *phonetic* knowledge of a prosodic category beyond just an overall baseline or range of a given cue (e.g., F0). For instance, a "question" prosody is hardly a homogeneous category either in its form or its meaning (Gunlogson, 2003; Hedberg et al., 2014). It has been argued that a question with declarative syntax (e.g., "It's raining?") is typically associated with a larger degree of pitch excursion compared to one with interrogative syntax (e.g., "Is it raining?") even when they are deemed equivocal in their phonological properties (Haan, 2001). Such systematic, phonetic variations may mean that not only do listeners need to adapt to talkers' base vocal pitch and pitch range, but they may also need to tune into *how these cross-talker variations might interact with*

syntactic contexts and subtle meaning differences associated with them (e.g., How do Talker A's declarative (vs. interrogative) questions differ from Talker B's?).

Functionally, optimal solutions for listeners can be affected by two motivations. On the one hand, listeners need to adapt and generalize the adaptation efficiently. Characteristics of a talker seen in one type of question should generalize to another (unseen) type of question, or else talker-adaptation would take too long to be practically useful for communication. On the other hand, listeners must adapt effectively and perhaps conservatively. Pitch movements observed for a declarative question, for example, might not directly predict how the same talker would produce an interrogative question. Such a balancing act has indeed been observed for adaptation to pronunciation variations in phonemes (e.g., vowels and consonants) and their positions in a word (Ades, 1974; Samuel, 1989) or lexical identity (Dahan & Mead, 2010). For example, stimulus features of a word-initial /d/ (e.g., *date*) and a word-final /d/ (e.g., *paid*) are typically distinct from each other, and exposure to one does not automatically produce talker-specific adaptation in the other (Samuel, 2020). The knowledge of fine-grained, context-dependent variations is thus thought to help reduce the risk of wholesale adaptation that is not warranted by natural distributions in the input. The current study asked if a similar context-sensitive conditioning would occur in prosodic adaptation.

2 Methods

2.1 Participants

180 participants were recruited through the online platform Amazon Mechanical Turk (https://www.mturk.com/). Five participants were excluded for providing a uniform response (either question or statement) for all tokens, leaving 175 participants for the analysis. Participants were all self-identified native speakers of American English and received monetary compensation ($4.00) for their participation.

2.1 Stimuli

The test stimuli were the same as those used in Xie, Buxó-Lugo, and Kurumada (2021, henceforth XBK2021), who demonstrated prosodic adaptation across one type of construction: "It's X-ing" (e.g., "It's cooking"). XBK2021 resynthesized 11-step continua (Figure 1, for details see Supplementary Information) between naturally produced falling (H* L-L%) and rising (H* L-H%) terminal pitch. The acoustic continua allowed for testing of listeners' categorization judgments of items into question vs. statement prosody. This was analogous to creating acoustic continuum between contrasting phonemes (e.g., with an unambiguous instance of "/p/each" at one end, and an unambiguous instance of "/b/each" at the other, see Norris et al. 2003; Clayards et al., 2008).

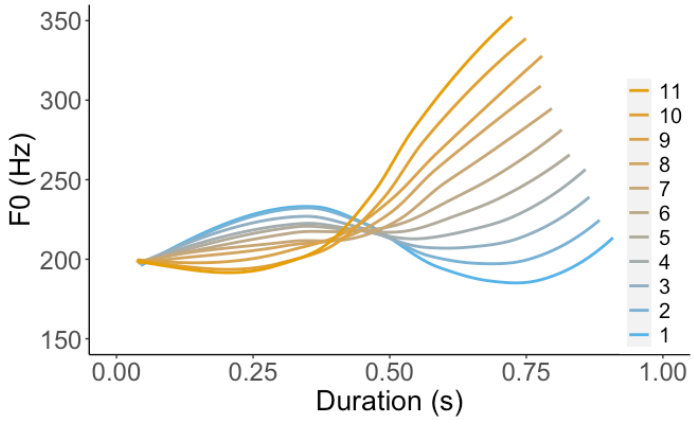

Figure 1. Pitch track and durational measures of test items ("It's cooking") used both in XBK2021 and in the current study (reprinted from XBK2021).

To create exposure stimuli, we recorded the same female native speaker of American English. The speaker produced the construction with the interrogative syntax "Is it X-ing" with the same five exposure verb types used in XBK2021 (i.e., booting, cooling, losing, moving, muting). These were spoken with the falling terminal pitch to form a basis of the same resynthesis procedure. The recordings were divided into three regions (i.e., is it | X | ing), and we imposed the f0 and duration measurements used in XBK2021 (taken from "It's moving") onto these new items to create 11 steps with Step 6 as the middle (stimuli available here: https://osf.io/hdftk/).

2.3 Procedure

The procedure was identical to XBK2021 except for the combinations of tokens heard during the exposure phase (Figure 2). Throughout the experiment, participants were asked to judge whether a given utterance was meant to be a question or a statement and provide their answer in a two-alternative forced choice task (2AFC) by clicking one of the two togglable buttons labeled as "Question" or "Statement." During the pre- and post-test, as in XBK2021, participants heard 22 tokens of "It's cooking" sampled uniformly along the 11-step continuum in a randomized order, with no feedback.

In exposure, participants were randomly assigned to one of the three conditions: *question-biasing*, *no-bias*, or *statement-biasing*. In all the conditions, they continued to answer 2AFC questions, now receiving feedback. In the no-bias condition, exposure tokens with statement feedback were sampled from Step 1 (the lowest pitch rise), and those with question feedback from Step 11 (the highest pitch rise) (Figure 1). In the statement-biasing condition, exposure tokens with statement feedback were sampled from Step 6, and those with question feedback from Step 11. In the question-biasing condition, exposure tokens with statement feedback were sampled from Step 1, and those with question feedback from Step 6. All participants heard the five exposure token types six times each: three times with statement feedback and three times with question feedback (30 total trials).

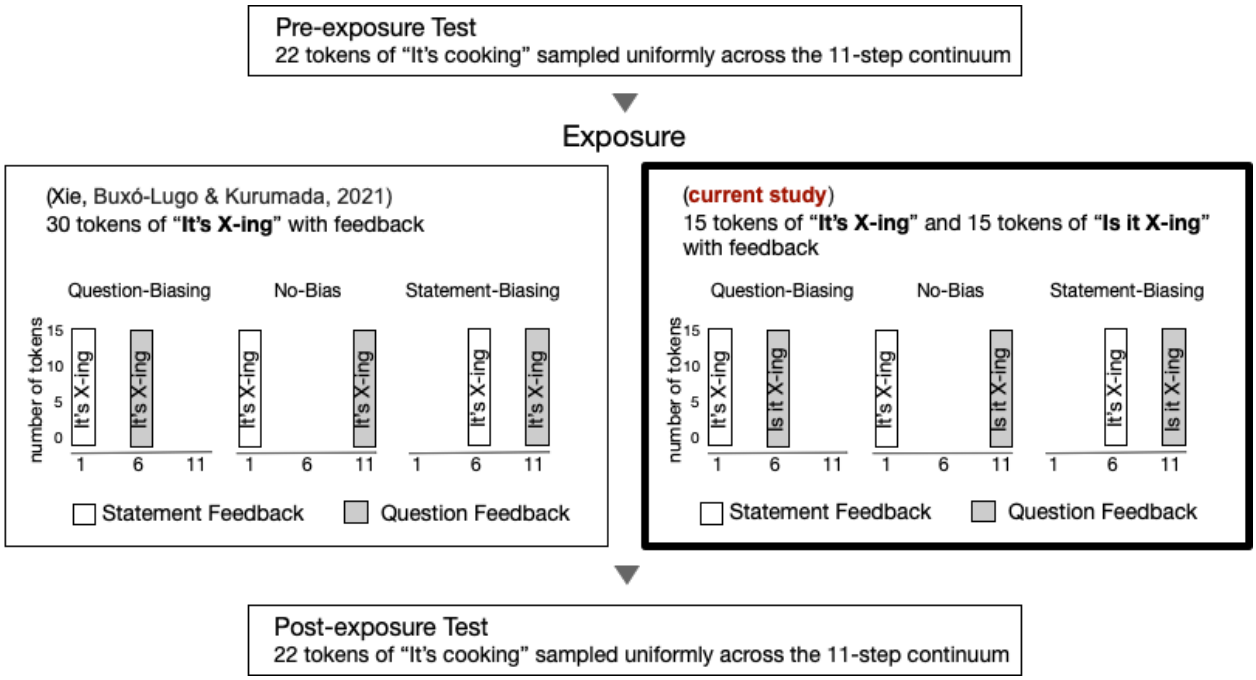

Figure 2. Schematic for the flow of the experiment in XBK2021 (left) and in the current study (right). In both, listeners repeatedly categorized tokens sampled uniformly from the 11-step continua in the pre- vs. post-exposure test blocks (22 tokens each). The only difference is the syntactic construction of a subset of the tokens used in the exposure phase. In the current study, all the exposure tokens associated with question feedback (gray bars) were with the interrogative syntax ("Is it X-ing"). Note that the interrogative syntax was used for tokens from either Step 6 or Step 11. It was never paired with the falling contour (i.e., Step 1).

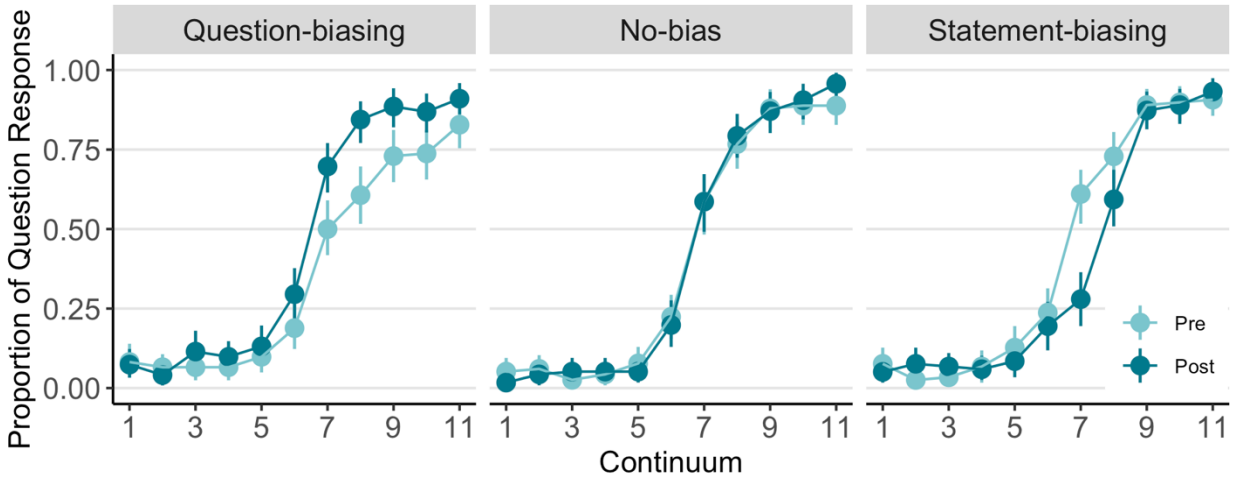

Figure 3. Proportions of question responses in the pre- vs. post-exposure test blocks across conditions in XBK2021. Error bars indicate bootstrapped 95% confidence intervals.

The sole difference between XBK2021 and the current study is that exposure tokens with question feedback were hosted in the interrogative syntax ("Is it X-ing") rather than in the declarative syntax ("It's X-ing") (Figure 2). Note that the interrogative syntax was used for Step 6 (i.e., ambiguous between falling and rising) or Step 11 (rising) depending on the condition but *never* for Step 1 (falling). If listeners were tracking pitch contours irrespective of the syntactic contexts, prosodic information to be learned from the exposure stimuli would be identical across the two studies, predicting replication of the results from XBK2021. Alternatively, if adaptation is conditioned on the syntactic context, it was expected only in the statement-biasing condition, where the ambiguous exposure tokens and test tokens shared the syntactic context. Limited or no adaptation was predicted for the question-biasing condition because the ambiguous tokens were heard with the different, interrogative syntax.

In both scenarios, the presence of contrasting syntactic contexts may prompt listeners to make an additional form-based inference (Grice, 1975) along the lines of "If the talker intended to ask a question, she should have used the interrogative syntax. 'It's X-ing' should therefore signal the statement meaning." Such an inference, although important in understanding pragmatic processing, is orthogonal to the prosodic adaptation we investigate here. We addressed this possibility by assessing the results of the no-bias condition. The potential form-based inference, if present, would trigger an overall bias towards statement responses at post-test.

3 Results

Results are summarized in Figure 4. To assess the effect of exposure for each of the conditions, we constructed a logistic mixed effects model using the R package BRMS (Bayesian Regression Models using Stan, Bürkner, 2019). The model had question responses as a dependent variable (coded as 1 = question or 0 = statement) and included block (post- vs pre-exposure, dummy-coded), condition (dummy-coded, with the no-bias condition as the comparison group), their interactions, and continuum step (centered). The random effects structure contained random slopes and intercepts per subject for block, condition, and continuum. Details of the model can be found in Supplementary Information.

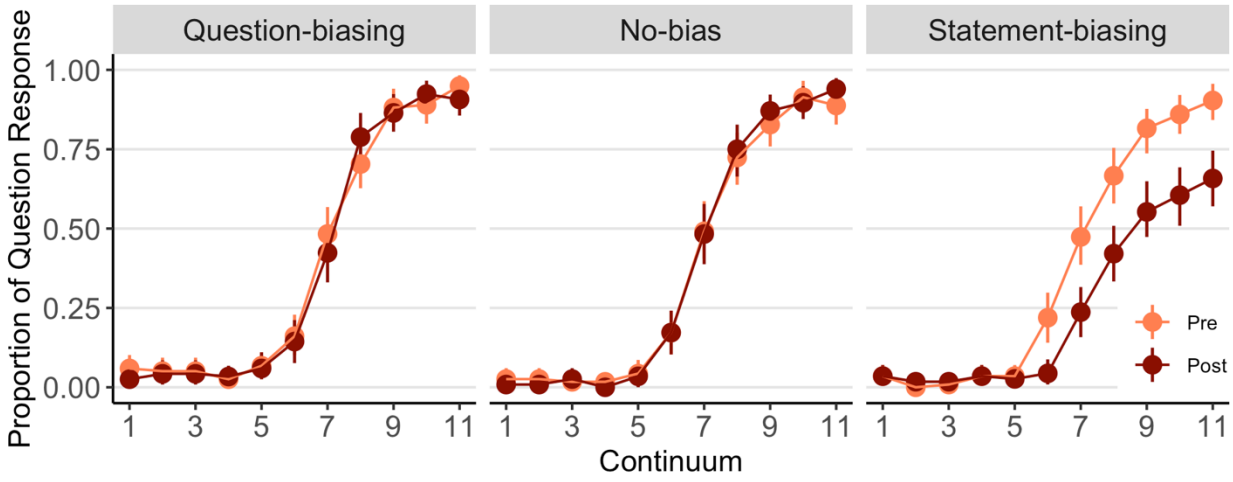

Figure 4. Proportions of question responses in the pre- vs. post-exposure test blocks across conditions in the current experiment. Error bars indicate bootstrapped 95% confidence intervals.

A model summary is provided in Table 1. BRMS, unlike frequentist models, estimates the posterior probability distribution given the observed data and specified priors. We report 95% credible intervals (CIs) to evaluate whether a given coefficient is meaningfully different from zero. As expected, the model found evidence for a main effect of the continuum step: Participants were more likely to provide a question response at higher steps along the continuum (β = 1.59, 95% CI = [1.44, 1.76]; Pr(β = 0) ≈ 0, Evidence Ratio = Infinite). Crucially, there was strong evidence for an interaction between the blocks and the statement-biasing condition (β = -2.15, 95% CI = [-2.98, -1.34]; Pr(β < 0) ≈ 1, Evidence Ratio = Infinite) while there was virtually no evidence for the interaction between the blocks and the question-biasing condition (β = -0.07, 95% CI = [-0.82, 0.73]; Pr(β > 0) ≈ 0.42, Evidence Ratio = 0.73). This difference supports the prediction that listeners condition their prosodic adaptation on syntactic contexts of heard items. In addition, the absence of the main effect of the block eliminated the possibility that the increased statement responses in the statement-biasing condition was simply due to a form-based inference between the two constructions.

Table 1: Summary of population-level (fixed) effects of the logistic mixed effects model using BRMS.

	Estimate	Estimated Error	95% CI
Intercept	-1.89	0.24	[-2.38, -1.41]
Block (Post-test vs Pre-test)	0.07	0.28	[-0.49, 0.63]
Question-biasing vs. No-bias	-0.10	0.35	[-0.79, 0.58]
Statement-biasing vs. No-bias	0.26	0.35	[-0.42, 0.96]
Continuum	1.59	0.08	[1.44, 1.76]
Block * Question-biasing	-0.07	0.40	[-0.82, 0.73]
Block * Statement-biasing	-2.15	0.42	[-2.98, -1.34]

To directly compare the results of the current experiment against those from XBK2021, we conducted a post-hoc analysis over a combined dataset using the same model specifications as described above (Supplementary Information). Figure 5 compares the posterior probability distributions for the critical two-way interactions across the experiments. The posterior distributions of coefficients for the block * statement-biasing interactions (in log-odds) were reasonably far away from 0 in both experiments; participants were significantly less likely to respond "question" after the exposure phase. In contrast, the effect of block * question-biasing condition was virtually nonexistent, unlike in XBK2021. Finally, posterior distributions of the three-way interaction between the block * question-biasing condition * experiment (XBK2021

vs. Current) was also meaningfully different from zero. This corroborates the idea that adaptation was blocked when the exposure and test tokens were associated with distinct syntactic contexts.

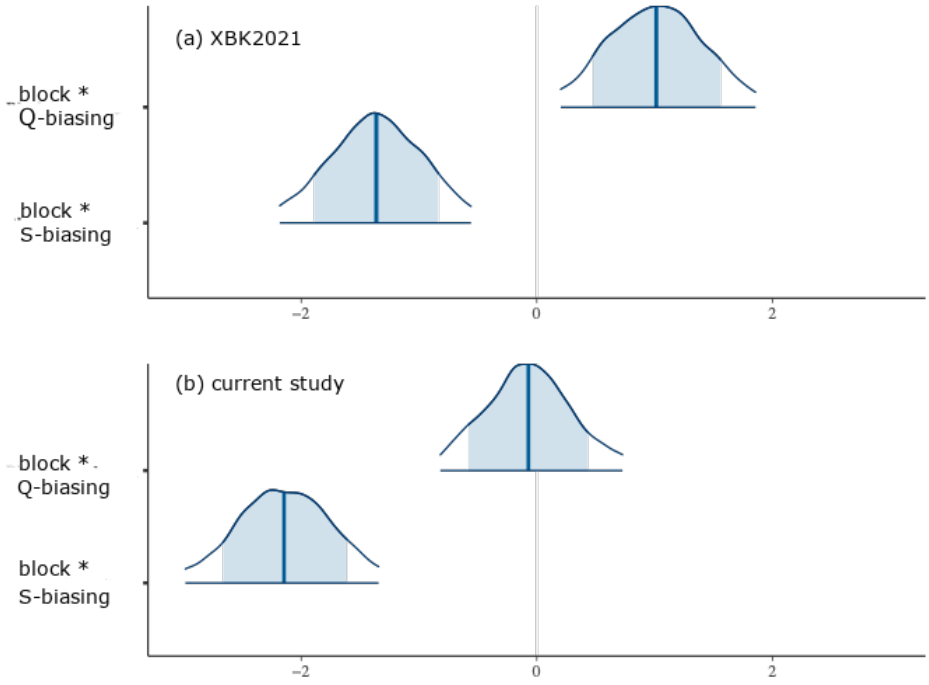

Figure 5. Posterior distributions of the two-way interaction terms in (a) XBK2021 and (b) the current study. Q-biasing = question-biasing; S-biasing = statement-biasing. The X-axis represents log-odds. Shaded areas and cut-off points of the density curves indicate 80% and 95% Highest Posterior Density Intervals (HPDIs), respectively.

4 Discussion

The importance of talker-normalized pitch and speech rate for prosodic processing is intuitive and has garnered much attention. Here, we examined the hypothesis that listeners' knowledge of prosodic variability includes fine-grained phonetic variations conditioned on their syntactic contexts. Between Xie et al. (2021) and the current study, the prosodically identical exposure tokens — "It's X-ing" vs. "Is it X-ing"— yielded strikingly distinct effects on listeners' post-test categorization judgments. Critically, participants who heard the ambiguous exposure tokens with the interrogative syntax did not alter their judgments on tokens with the declarative syntax before and after the exposure. If adaptation relied solely on adjusting estimates of the talker's overall base vocal pitch and/or pitch range, this result would not be predictable (for novel, corroborating evidence that prosodic adaptation requires a linguistically intelligible input beyond a talker's pitch information, see Bosker (2021)).

The absence of adaptation in the question-biasing condition and the (numerically) enhanced adaptation in the statement-biasing condition can reveal interesting dynamics present in speech adaptation. One possibility is that the patterns could be indicative of error-based learning as a necessary component of speech adaptation (e.g., Dell & Chang, 2013; Fine &

Jaeger, 2013). A prosodically ambiguous exposure token in the question-biasing condition was consistently marked as a question via the interrogative syntax (i.e., the subject-auxiliary inversion). Listeners could therefore rely solely on the syntactic marking to answer the 2AFC questions, which might make prosodic adaptation less consequential for comprehension. Although possible, this account runs counter to prevailing theories of speech perception, where recalibration of speech categories is considered largely implicit and automatic. For example, exemplar-based accounts assume that each new exemplar is indiscriminately stored. Adaptation results from subsequent input being categorized with reference to all previously experienced exemplars (e.g., Hay & Drager, 2007; Johnson, 2006; Pierrehumbert, 2001). We know of no account that indicates that speech adaptation occurs exclusively when it is consequential for lexical/meaning disambiguation. If anything, labeling (or disambiguating) information from a lexical or syntactic context (e.g., "croco[?]ile" for the /d/ category) is thought to guide phonetic adaptation rather than block it (e.g., Norris et al., (2003); Kraljic & Samuel (2005)). For this reason, the availability of the syntactic marking of question-hood alone may be insufficient to account for the discrepancy between the current results and those from XBK2021.

An alternative, though not mutually exclusive, explanation would involve the role of listeners' fine-grained, phonetic expectations about realizations of declarative vs. interrogative questions, against which the current exposure input was processed. As mentioned in the introduction, declarative questions in English tend to show an overall more pronounced level of terminal pitch rise than interrogative questions (Haan, 2001; Grabe, 2002). Note that the acoustic continua used in the current study were created originally from values taken from declarative question. This could mean that a token of the interrogative question that was meant to be prosodically "ambiguous" (Step 6) may have had a terminal pitch rise large enough for an interrogative question. If so, listeners in the question-biasing condition might not have had much evidence that the talker's uses of prosody were in any way deviant from what is normally expected, and hence no adaptation was needed. According to the same logic, in the statement-biasing condition, the pitch rise experienced with the interrogative syntax (Step 11) during exposure might have been more extreme than what would be generally expected. Participants might have then inferred that post-exposure test tokens must have an even more pronounced rise for them to count as (declarative) questions. This correctly predicts the significant adaptation seen in the statement-biasing condition.

An empirical test of this hypothesis requires assessments of natural production data, detailing whether and how much prosodic variability is observed for different syntactic contexts (e.g., declarative vs. interrogative questions). Such production data can be used to approximate listeners' expectations at the outset of an experiment. Once made available, a large-scale corpus of this kind will help address other outstanding questions. Chief among them is which syntactic constructions serve as conditioning contexts for prosodic adaptation. Do listeners store all possible syntax-prosody combinations? Or do they represent only those that *typically* vary within/across talkers? Recent development of computational approaches has made it possible to quantify the amount of variability associated with a given context, be it linguistic (e.g., lexico-syntactic) or socio-indexical (Chodroff & Wilson, 2019; Kleinschmidt, 2019). One can estimate the amount of perceptual benefit a listener could, in principle, gain if the variability is effectively conditioned on context (e.g., How much more accurately can one distinguish a question from a statement if the identity of a talker is known vs. unknown? Xie, Buxó-Lugo & Kurumada, 2021). Extending these approaches will facilitate investigations into how listeners may represent and weight multiple sources of variability to achieve reliable comprehension of speech prosody.

In summary, listeners tune into phonetic details of prosody as they adapt to individual talkers. This finding resonates with existing linguistic and psycholinguistic theories, wherein listeners are believed to leverage prosodic details to distinguish between subtle shades of meaning. What we called declarative questions in this study, for instance, can be further sub-divided according to their form and meaning (Jeong & Potts, 2016). However, few accounts have been given so far to explain how listeners detect these subtle differences in the sea of talker-variability in human speech. The current finding suggests that listeners can achieve this by flexibly adapting their prosodic processing in a manner attuned to variations expected across linguistic (e.g., syntactic) contexts. This way of reasoning is analogous to position-dependent adaptation of phonemes (i.e., word initial, medial, and final). Overlaying variations stemming from talkers and talker groups can be best learned if the perceived stimulus features are normalized against the *a priori* expectations given a position effect. We conclude that, also in prosodic processing, listeners may condition their expectations according to lexical and syntactic contexts, creating a basis on which to accommodate cross-talker variability.

Open Practices Statement

The data, stimuli, and supplementary analyses for the experiment are available at https://osf.io/hdftk/. The experiment was not pre-registered.

References

Ades, A. (1974). How phonetic is selective adaptation? Experiments on syllable position and vowel environment. *Perception & Psychophysics, 16*, 61–66.

Arvaniti, A. (2011). The representation of intonation. In *Blackwell Companion to Phonology* (pp. 757–780). Oxford, UK: Oxford University Press.

Beckman, M. E. (1995) Local shapes and global trends. *Proceedings International Conference of Phonetic Sciences*, Stockholm, Vol. II, (pp. 100-107).

Bishop, J., & Keating, P. (2012). Perception of pitch location within a speaker's range: Fundamental frequency, voice quality and speaker sex, *The Journal of the Acoustical Society of America 2012, 132* (2), 1100-1112.

Bolinger, D. (1964). Intonation as a universal. In *Proceedings at Linguistics IX* (pp. 833–844). The Hague: Mouton.

Bosker, R. H. (2021). Evidence for selective adaptation and recalibration in the perception of lexical stress. *Language and Speech*.

Breen, M., Kurumada. C., Wagner, M., Watson, D., & Yu, K. (2018). Introducing prosodic variability. *Laboratory Phonology: Journal of the Association for Laboratory Phonology, 9*(1), 5. DOI: http://doi.org/10.5334/labphon.142

Bürkner, P.-C. (2019). Bayesian item response modelling in R with BRMS and Stan. Retrieved from https://arxiv.org/abs/1905.09501

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.

Cole, J. (2015). Prosody in context: A review. *Language, Cognition and Neuroscience*, *30*(1–2),

1–31.
Cutler, A. (2015). *Native listening: Language experience and the recognition of spoken words*. MIT Press.
Dahan, D. (2015). Prosody and language comprehension. *Wiley Interdisciplinary Reviews: Cognitive Science*, *6*(5), 441–452. https://doi.org/10.1002/wcs.1355
Dahan D., & Mead, R. L. (2010). Context-conditioned generalization in adaptation to distorted speech. J Exp Psychol Hum Percept Perform. 36(3): 704-28. doi: 10.1037/a0017449. PMID: 20515199.
Dell, G. S., & Chang, F. (2013). The P-chain: relating sentence production and its disorders to comprehension and acquisition. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, *369*(1634), 20120394. https://doi.org/10.1098/rstb.2012.0394
D'Imperio, M., Cangemi, F., & Grice, M. (2016). Introducing advancing prosodic transcription. *Laboratory Phonology: Journal of the Association for Laboratory Phonology, 7*(1). https://doi.org/10.5334/labphon.32
Dilley, L. C., & Pitt, M. A. (2010). Altering context speech rate can cause words to appear or disappear. *Psychological Science*, *21*(11), 1664–1670.
Fine, A.B., Jaeger T.F., (2013). Evidence for implicit learning in syntactic comprehension. *Cognitive Science, 37*(3):578-91. Doi: 10.1111/cogs.12022.
Grabe, E. (2002). Variation adds to prosodic typology. *Speech Prosody 2002*, 127–132.
Grice, P. (1975). Logic and Conversation. In P. Cole & J. Morgan (Eds.), *Syntax and Semantics, vol.3* (pp. 41–58). New York: Academic Press.
Gunlogson, C. (2003). *True to form: Rising and falling declaratives as questions in English*. New York: Routledge.
Gussenhoven, C., & Rietveld, T. (1996). On the speaker-dependence of the perceived prominence of F0 peaks. *Journal of Phonetics, 26*, 371-380.
Hawkins, S. (2003). Roles and representations of systematic fine phonetic detail in speech understanding. *Journal of Phonetics, 31*. 373-405. 10.1016/j.wocn.2003.09.006.
Haan, J. (2001). *Speaking of questions: An exploration of Dutch question intonation*. *LOT Dissertation Series* (Vol. 52).
Hay, J., & Drager, K. (2007). Sociophonetics. *Annual Review of Anthropology*, *36*(1), 89–103. https://doi.org/10.1146/annurev.anthro.34.081804.120633
Hedberg, N., Sosa, J. M., & Görgülü, E. (2017). The meaning of intonation in yes-no questions in American English: A corpus study. *Corpus Linguistics and Linguistic Theory*, *13*(2), 321–368.
Ito, K., Turnbull, R., & Speer, S. R. (2017). Allophonic tunes of contrast: Lab and spontaneous speech lead to equivalent fixation responses in museum visitors. *Laboratory Phonology*, *8*(1). https://doi.org/http://doi.org/10.5334/labphon.86
Jeong, S., & Potts, C. (2016). Intonational sentence-type conventions for perlocutionary effects: An experimental investigation. *Semantics and Linguistic Theory*, *26*, 1. https://doi.org/10.3765/salt.v26i0.3787
Johnson, K. (2006). Resonance in an exemplar-based lexicon: The emergence of social identity and phonology. *Journal of Phonetics*, *34*(4), 485–499.
Kleinschmidt, D. F. (2019). Structure in talker variability: How much is there and how much can it help? *Language, Cognition and Neuroscience*, *34*(1), 43–68. https://doi.org/10.1080/23273798.2018.1500698
Kleinschmidt, D. F., Weatherholtz, K., & Florian Jaeger, T. (2018). Sociolinguistic perception as

inference under uncertainty. *Topics in Cognitive Science*, *10*(4), 818–834. https://doi.org/10.1111/tops.12331

Kurumada, C., Brown, M., & Tanenhaus, M. K. (2017). Effects of distributional information on categorization of prosodic contours. *Psychological Bulletin and Review*. https://doi.org/10.3758/s13423-017-1332-6

Ladd, D. R. (2008). *Intonational phonology* (2nd ed.). Cambridge University Press.

Lee, C.-Y. (2009). Identifying isolated, multispeaker Mandarin tones from brief acoustic input: A perceptual and acoustic study. *Journal of the Acoustical Society of America*, *125*, 1125–1137.

Nakamura, C., Harris, J., & Jun, S.-A. (2019). Listeners' beliefs about the speaker and adaptation to the deviant use of prosody. In *Poster presented at the 32nd annual CUNY Conference on Human Sentence Processing, Boulder, CO*.

Norris, D., McQueen, J. M., & Cutler, A. (2003). Perceptual learning in speech. *Cognitive Psychology*, *47*(2), 204–238.

Patel, R., & Brayton, J. T. (2009). Identifying prosodic contrasts in utterances produced by 4, 7, and 11 year old children. *Journal of Speech, Language, and Hearing Research*, (June), 790–801.

Patel, R., Niziolek, C., Reilly, K., & Guenther, F. H. (2011). Prosodic adaptations to pitch perturbation in running speech. *Journal of speech, language, and hearing research*, *54*(4), 1051–1059. https://doi.org/10.1044/1092-4388(2010/10-0162)

Pierrehumbert, J. B. (2001). Exemplar dynamics: Word frequency, lenition and contrast. In J. Bybee & P. Hopper (Eds.), *In Frequency and the Emergence of Linguistic Structure* (pp. 137–157). John Benjamins.

Pierrehumbert, J., & Hirschberg, J. (1990). The meaning of intonational contours in the interpretation of discourse. In P. R. Cohen, J. Morgan, & M. E. Pollack (Eds.), *Intentions in communication* (pp. 271–311).

Saindon, M. R., Trehub, S. E., Schellenberg, E. G., & van Lieshout, P. (2017). When is a question a question for children and adults?, *Language Learning and Development, 13*(3), 274-285, DOI: 10.1080/15475441.2016.1252681

Samuel, A.G. (1989). Insights from a failure of selective adaptation: Syllable-initial and syllable-final consonants are different. *Perception & Psychophysics, 45*(6), 485-493. DOI: 10.3758/BF03208055

Samuel, A. G. (2020). Psycholinguists should resist the allure of linguistic units as perceptual units. *Journal of Memory and Language*, *111*, 104070–104070. https://doi.org/10.1016/j.jml.2019.104070

Schweitzer, K., Walsh, M., Calhoun, S., Schütze, H., Möbius, B., Schweitzer, A., & Dogil, G. (2015). Exploring the relationship between intonation and the lexicon: Evidence for lexicalized storage of intonation. *Speech Communication, 66,* 65-81.

Tang, C., Hamilton, L. S., & Chang, E. F. (2017). Intonational speech prosody encoding in the human auditory cortex. *Science*, *801*, 797–801.

Woodard, K., Plate, R.C., Morningstar, M., Wood, A., & Pollak, S. D., (2021). Categorization of vocal emotion cues depends on distributions of input. *Affective Science* (2021). https://doi.org/10.1007/s42761-021-00038-w

Xie, X., Buxó-Lugo, A., & Kurumada, C. (2021). Encoding and decoding of meaning through structured variability in speech prosody. *Cognition*, *221*. https://doi.org/10.1016/j.cognition.2021.104619

Supplementary Information

S.1 Statistical Analysis of the Data

To analyze the data, we used Bayesian hierarchical models using the Stan (BRMS) modelling language (Carpenter et al., 2017) and the R package brms (Bürkner, 2019). All data tables and R scripts are available here: https://osf.io/hdftk/. Here, we describe the statistical analyses we conducted and assumptions applied in the analyses. We acknowledge that we are in debt to careful documentations provided in prior studies, especially Roettger and Baer-Henney (2018). We chose BRMS over a canonical frequentist model for the current analysis for two reasons.

First, within the constraints of the current design, the Bayesian framework allowed us to construct a model with the maximal random effect structure justified by the data. We have constructed linear mixed effects models using the lme4 package in R (Bates, Mächler, Bolker, & Walker, 2015) with the random intercepts and slopes (as recommended by Schielzeth & Forstmeier (2008); Barr, Levy, Scheepers, & Tily (2013); Bates, Kliegl, Vasishth, & Baayen (2015)), but those models failed to converge. BRMS allowed us to fit the maximal random effect structure, which is considered more conservative.

Second, the Bayesian framework enables us to quantitatively estimate the likelihood of a particular hypothesis given the observed data. In other words, this modeling method can quantify our uncertainty about the parameters of interest, "which frees us from committing to hard cut-off points for statistical significance (such as the arbitrary .05 alpha level)" (Roettger & Baer-Henney, 2018). This was better suited than a frequentist model to our goal of comparing the results of the two experiments to infer *an extent to which* the current design induced results different from those in XBK2021.

In analyzing the current experiment, we fit hierarchical regression models to log-odds of binomial responses (question vs. statement) predicted by Conditions (question-biasing, no-bias, statement-biasing, dummy coded as the no-bias condition as the baseline), Blocks (pre, post, dummy-coded), their interactions, and Continuum (1-11, centered). The models included a random-effect structure that included by-subject random slopes and intercepts. Following recommendations from the statistical literature (Gelman et al., 2013; Bürkner, 2019), we used the default priors in BRMS, which are meant to be weakly informative. However, it is worth noting that specifying priors based on the results from XBK2021 did not meaningfully change the pattern of results.

The model was fit via Stan's No-U-Turn sampler — a Hamiltonian Monte Carlo Method with many computationally desirable properties (Monnahan, Thorson, & Branch, 2017). Four sampling chains with 2000 iterations were run for each model, with a warm-up period of 1000 iterations. The chains mixed well (e.g., all R-hats ~ 1.001 or closer to 1), and there were no divergent transitions after warm-up.

We report 95% credible intervals (CIs). A 95% credible interval "demarcates the range of values that comprise 95% of the probability mass of our posterior beliefs" (Roettger & Baer-Henney, 2018). It is generally accepted that there is compelling evidence for an effect if 0 is not included in the 95% CI. As part of the model summary, we report an evidence ratio as derived by the "hypothesis" function included in BRMS, as well as an estimated probability of the draws

from the posterior distributions of the critical interactions to be larger or smaller than 0 in an expected direction, that is $Pr(\beta < 0)$ or $Pr(\beta > 0)$. This estimate provides supportive evidence for the alternative hypothesis, as opposed to the null hypothesis, if it is close to 1.

The model output is summarized in Table 1 in the manuscript. Here we include a corresponding visual representation of posterior uncertainty distributions for the fixed effects in Figure s1. As we mentioned above, a given fixed effect can be considered "significant" (meaningfully distinguishing the alternative from the null hypothesis) if a distribution is away from 0 by a sufficient margin. As can be seen in Figure s1 above, the distributions for the main effects of Block and Conditions (question-biasing / statement-biasing) cross 0, hence those effects are unlikely to be significant. On the other hand, the main effect of Continuum is, as expected, highly likely to be significant. That is, participants were more likely to provide the question response for the items sampled from a region closer to the higher end of the continuum.

Distributions of the two-way interaction terms, Block * question-biasing / statement-biasing Conditions, exhibited different patterns. The interaction between Block and statement-biasing condition is significantly away from 0, suggesting that participants in the statement-biasing condition were less likely to provide the question response in the post-exposure phase. In contrast, the interaction between Block and question-biasing condition crosses 0. This is a pattern expected under the prediction that intonation adaptation is conditioned on syntactic constructions: In the question-biasing condition where the question meaning was expressed with distinct constructions between the exposure and test stimuli phases, the adaptation was unlikely to occur.

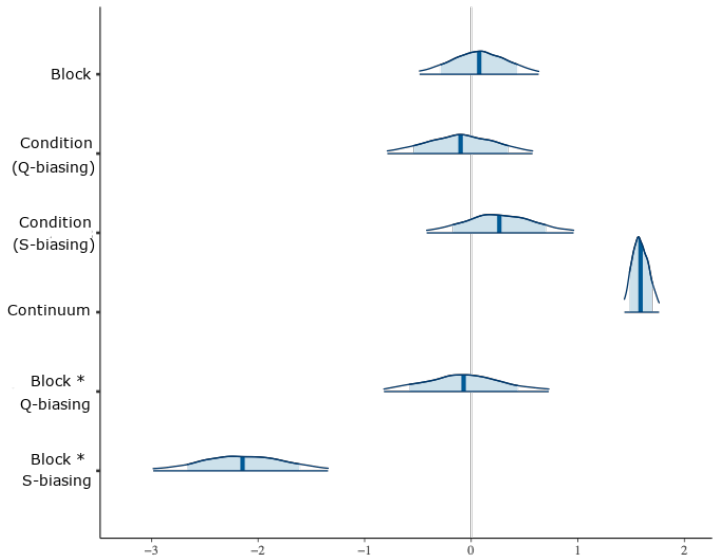

Figure s1. Posterior uncertainty distributions of the fixed effects of the BRMS model. The X-axis represents log-odds. S-biasing = Statement-biasing Condition; Q-biasing = Question-biasing Condition. Shaded areas indicate 80% Highest Posterior Density Intervals (HPDIs), and the distributions are cut off at 95% HPDIs.

S.2 Statistical Analysis of the Data from XBK 2021

The data from XBK2021 were analyzed with the same exact model structure and coding schemes as those described for the current study. Packages and tools used were identical to those in the

current study to allow for comparisons of the model outcomes from the two studies. The summary of the fixed effect in the model of the XBK2021 data is presented in Table s1 below. Overall, the results were similar to those from the present study. A key difference was that, whereas the present study found no evidence for the presence of a Block * question-biasing interaction, XBK2021 did find evidence for this interaction in the expected direction (β = 1.01, 95%CI = [.20, 1.86]; Pr(β > 0) = 1). Participants in the question-biasing condition in XBK2021, unlike those in the current experiment, were more likely to provide the question response during the post-exposure block.

Table s1. Summary of population-level (fixed) effects of the logistic mixed effects model using BRMS run on the combined dataset from XBK2021 and the current study.

	Estimate	Estimated Error	95% CI
Intercept	-1.62	0.28	[-2.18, -1.08]
Block (Post-exposure vs Pre-exposure)	0.06	0.30	[-0.53, 0.65]
Question-Biasing vs No-bias	-0.35	0.41	[-1.15, 0.45]
Statement-Biasing vs No-bias	0.17	0.38	[-0.58, 0.94]
Continuum	1.70	0.11	[1.50, 1.91]
Block * Question-Biasing	1.01	0.42	**[0.20, 1.86]**
Block * Statement-Biasing	-1.36	0.41	[-2.18, -0.56]

S.3 Statistical Analysis of the Combined Dataset

To directly compare the results of the two experiments, we conducted an analysis on a combined dataset including the data from both XBK2021 and the current experiment. The model specifications are identical to what we described under the Results section of the manuscript and Section 1 above except for the Experiment factor (XBK2021 vs. Current, dummy-coded). This was included as a fixed effect with an interaction term with Block and Condition, and in the random effects structure as independent slopes and intercepts per subject. By including the interaction between Experiment and the crucial Block * Condition interaction, we can test whether the effects of Block * Condition were significantly different between XBK2021 and the current experiment.

The summary of the fixed effects of the model is given in Table s2 below. We found evidence for a three-way interaction between Block, Condition, and Experiment for the question-biasing conditions (β = -1.18, 95%CI = [-1.84, -0.54]; Pr(β < 0) = 1). Participants in the current experiment, as compared to those in XBK2021, were significantly *less* likely to provide the question response after the Exposure phase. This supports the conclusion that the degrees of adaptation seen in the question-biasing condition differed across these experiments.

We note also that the three-way interaction was significant for the S-biasing (β = -0.85, 95%CI = [-1.47, -0.23]; Pr(β < 0) = 1). This likely reflects the strong bias we observed at the higher region along the continuum (Figure 4). Compared to those in XBK2021, participants in the statement-biasing condition in the current experiment were more likely to provide the

statement response in the post-exposure test. As we discuss in the Discussion, this could be due to the fact that the exposure tokens in this condition strongly biased the listener to expect a prominent rise for a declarative question. In Exposure, listeners in this condition heard "It's X-ing" tokens from Step 6 as associated with the statement meaning and "Is it X-ing" tokens from Step 11. This pattern of exposure supports the following observations:

1) The "It's-Xing" tokens were associated with more terminal rise than expected.
2) The tokens of "Is it X-ing" have a more prominent terminal pitch rise than would be generally expected in productions. (Recall that these tokens were resynthesized based on the intonational features of a declarative question.)

Based on these observations, listeners might have concluded that this particular talker tends to produce more extreme terminal pitch rise than normally expected. This would support an inference that a *declarative* question produced by the same talker would have even more extreme rise. If this is the case, we can straightforwardly explain the increased statement responses in the post-exposure test. Though the validity of such inferences needs to be directly verified in future studies, we conclude here that listeners' judgments as observed in the post-exposure test are predictable based on the types of intonation variations typically associated with the two distinct constructions used in the experiment.

Table s2. Summary of population-level (fixed) effects of the logistic mixed effects model using BRMS run on the combined dataset from XBK2021 and the current study.

	Estimate	Estimated Error	95% CI
Intercept	-1.27	0.18	[-1.64, -0.91]
Block (Post-Exposure vs Pre-Exposure)	0.02	0.17	[-0.32, 0.34]
Question-Biasing vs No-bias	-0.39	0.25	[-0.88, 0.11]
Statement-Biasing vs No-bias	0.13	0.25	[-0.36, 0.63]
Experiment (Current vs. XBK2021)	-0.30	0.26	[-0.81, 0.22]
Continuum	1.24	0.05	[1.14, 1.35]
Block * Question-Biasing	0.96	0.23	[0.51, 1.40]
Block * Statement-Biasing	-0.68	0.23	[-1.13, -0.24]
Block * Experiment	-0.02	0.22	[-0.44, 0.40]
Question-Biasing * Experiment	0.54	0.37	[-0.18, 1.28]
Statement-Biasing * Experiment	-0.29	0.38	[-1.03, 0.45]
Block * Question-Biasing * Experiment	-1.18	0.33	[-1.84, -0.54]
Block * Statement-Biasing * Experiment	-0.85	0.32	[-1.47, -0.23]

References

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278. https://doi.org/10.1016/j.jml.2012.11.001

Bates, D., Kliegl, R., Vasishth, S., & Baayen, H. (2015). Parsimonious mixed models. Retrieved from http://arxiv.org/abs/1506.04967

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Bürkner, P.-C. (2019). Bayesian Item Response Modelling in R with brms and Stan. Retrieved from https://arxiv.org/abs/1905.09501

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., … Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software; Vol 1, Issue 1 (2017)*. Retrieved from http://dx.doi.org/10.18637/jss.v076.i01

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis* (3rd ed.). Chapman and Hall/CRC.

Monnahan, C. C., Thorson, J. T., & Branch, T. A. (2017). Faster estimation of Bayesian models in ecology using Hamiltonian Monte Carlo. *Methods in Ecology and Evolution*, *8*(3), 339–348. https://doi.org/10.1111/2041-210X.12681

Roettger, T. B., & Baer-henney, D. (2018). Toward a replication culture: Speech production research in the classroom. *PsyArXiv*, 1–26. https://doi.org/10.17605/OSF.IO/9KYWF

Schielzeth, H., & Forstmeier, W. (2008). Conclusions beyond support: Overconfident estimates in mixed models. *Behavioral Ecology*, *20*(2), 416–420. https://doi.org/10.1093/beheco/arn145