NORTHWESTERN UNIVERSITY

Structured Variation in Intonational Form and Interpretation in American English

A DISSERTATION

SUBMITTED TO THE GRADUATE SCHOOL
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

for the degree

DOCTOR OF PHILOSOPHY

Linguistics

By

Thomas Sostarics

EVANSTON, ILLINOIS

June 2025

ABSTRACT

In Mainstream American English (MAE), phrasal intonation conveys distinctions in pragmatic meaning. Of particular importance are the intonational patterns that occur at the end of the intonational phrase: the nuclear tune. The dominant Autosegmental-Metrical (AM) theory of intonational phonology makes explicit predictions about the contrastive intonational *forms* in MAE. Findings that tune-level contrasts are associated with robust meaning distinctions in intonational *function* would be important evidence for the intonational inventory predicted by AM theory. Ideally, investigation of form and function would be a joint enterprise, but in practice the two perspectives are rather disjoint from each other. Ongoing debates about category-level distinctions among intonational features do not often inform work on intonational function. Conversely, ongoing debates regarding how best to characterize the meaning contribution of specific intonational features do not always inform work on intonational form.

Two phenomena stand out in the intonational meaning literature as having received ample investigation in terms of their function in MAE: rising declaratives and the rise-fall-rise tune. Investigations of these phenomena are largely separate from one another but share a common problem regarding how differences in the phonetic expression, or potentially the phonological status, of an intonational pattern may relate to distinctions in meaning. For rising declaratives, it has been observed that there is a distinction in function for shallow versus steep rises, but relating "shallow" and "steep" to proposed phonological distinctions is unclear. On the one hand, there may be a single, phonologically defined rising tune which varies in its phonetic implementation between relatively stable endpoints of shallow and steep; on the other hand, shallow and steep rising tunes may be phonologically distinct. For the rise-fall-rise (RFR) tune, researchers vary in whether they take variation in the initial rise (corresponding to the pitch accent in an AM formulation) to be phonologically contrastive, resulting in three distinct RFR-shaped tunes, or the same observed patterns may be reflective of paralinguistic variation with all variants sharing a common core to their meaning contribution. The view that motivates this thesis is that both rising declaratives and work

on RFR may benefit from providing closer attention to the intonational form.

This dissertation presents a series of perception and comprehension experiments tackling the question of between- versus within-category variation in rising declaratives and RFR in MAE. In particular, the present work provides an in-depth perceptual investigation of the range of phonetic variation possible for rises (vis-à-vis falls) and RFR. The results of Part 1, focusing on rising declaratives, suggest that variation in the region of the nuclear tune associated with the pitch accent does not play a robust role in conveying meaning related to speaker inquisitiveness versus assertiveness. Instead, variation in the ending F0 target is shown to be the strongest cue related to this meaning dimension for rises and falls. The results of Part 2, focusing on RFR, show that of the three phonologically distinct tunes with the RFR shape, all contribute similarly to pragmatic meaning in that all increase the likelihood of scalar inference computation. But other results show different patterns for these three RFR tunes in online processing in a cross-modal lexical decision paradigm. Specifically, RFR-shaped tunes using an H* pitch accent show increased facilitation of higher scalar alternatives while those using an L*+H pitch accent show less facilitation. These results are taken as evidence for a broad RFR class of tunes with meaningful within-category variation related to pitch range, which covaries with the phonological choice of pitch accent. Overall, this work provides an in-depth look at the contrast between inquisitive and assertive interpretations for rising declaratives, and also provides a wealth of novel evidence on the interpretation and processing of RFR in the context of scalar inference.

ACKNOWLEDGMENTS

Writing a dissertation is not easy, and as anyone who has talked to me in the past two years knows, I am **not** fond of writing. Despite this, I need to write just a little bit more to properly convey my gratitude to a number of people.

I must thank my committee—Jennifer Cole, Eszter Ronai, and Gregory Ward—for their encouragement to not only make this work better but to help me see value in this work even when I was sick of it. In particular, I could not have asked for a better advisor than Jennifer to push me to be a better scientist and prosody researcher. Whenever I found myself lost among the trees, she has always helped me take a step back to better see, understand, and appreciate the whole forest. Gregory has always supported my interest in intonational meaning, and I have always appreciated his expertise and insightful comments. I am also very thankful for Eszter's support in building up the psycholinguistic rigor in this work, as she has always pushed me to consider this work in the context of complementary topics and debates. Finally, thanks to Duane Watson, who served on my prospectus committee and provided sage advice about the priming task in this work; to share with future readers: Lexical decision is a **huge pain**.

My growth as a scholar has been influenced by a broader community of people who have provided feedback, support, and good conversations. In particular, I have gained much from conversations with Ann Bradlow and Mike Tabatowski, who both have a talent for identifying the fundamental principles that lie at the heart of problems related to both phonetics and meaning, respectively. I am also indebted to Ming Xiang, who offered my start as a fledgling experimentalist in her lab and has continued to offer me support and advice ever since. I have also benefited greatly from scholarly conversations with Khalil Iskarous, Daniel Goodhue, Bodo Winter, Kathryn Franich, Jason Bishop, and Jeremy Steffman. Finally, I am thankful for the camaraderie from all the conference pals I have made through the Midphon and international speech prosody communities.

I am immensely thankful for the friends I've made in the Linguistics communities at both

Northwestern University and The University of Chicago. In particular, I am indebted to my cohortmates Kate Sandberg, Chantal de Leon, and Anna Robinson, whose support has been the foundation of my time in graduate school. To Kate especially, I could always count on you to help me see the silver lining in all things (and to record some stimuli here and there). I am also thankful to my fellow ProSD Lab members past and present as well as María Gavino, Wesley Orth, Jaime Benheim, Shawn Foster, and others in the Northwestern Linguistics community. I am also thankful for the support from the UChicago linguistics community, especially Sanghee Kim, Daniel Lam, Lucas Fagen, and Aurora Martínez del Rio. I also want to thank the members of the former Aphasia Lab, especially Cynthia Thompson and Matt Walenski, who provided me valuable interdisciplinary opportunities at Northwestern.

Although many of my friends are also linguists, there are many others who aren't that have supported me just as much—even if they didn't really know exactly what I was doing. To Tim Johnson and Michelle Skinner in particular (and by extension, Cooper and Harper), thank you for always being there for me and for watching out for me. To Owen McNamara, Sam Dubensky, Mark Agrios, Guna Kondapaneni, Emily Lewis, JP Miller, Anlil Oraha, and Elaine Coppe: Thank you **all** for your continued support and check-ins over the years. A special thanks is in order for my best friend Carson Chauvin, whom I could always count on to keep me sane. Finally, thank you to my family, who have always given me an escape from the city.

LIST OF ABBREVIATIONS

AM	Autosegmental-Metrical
AP	Accentual Pitch
ARD	Assertive Rising Declarative
AUC	Area Under the [ROC] Curve
CI	Confidence Interval
CLM	Cumulative Link Model
ConfRD	Confirmative Rising Declarative
CPP	Cepstral Peak Prominence
CrI	Credible interval
CT	Contrastive Topic
CV	Cross Validation
disc	Discrimination parameter
ELPD	Expected Log Predictive Density
EP	Ending Pitch
ERB	Equivalent Rectangular Bandwidth
F0	Fundamental Frequency
GAMM	Generalized Additive Mixed Model
HF16	Husband & Ferreira (2016)
IC	Information Criterion
IncRD	Incredulous Rising Declarative
IP	Intonational Phrase
ip	Intermediate Phrase
IRD	Inquisitive Rising Declarative
LOO	Leave One Out
MAE	Mainstream American English
MFA	Montreal Forced Aligner
MI	Merely Inference
NA	Not Applicable
NG	Neo-Gricean
PD	Probability of Direction

PG Post-Gricean
PSOLA Pitch Synchronous Overlap and Add

QUD Question Under Discussion

RD Response Duration
RFR Rise-Fall-Rise
ROC Receiver Operator Characteristic
RT Reaction Time

sd Standard Deviation
SE Standard Error
SI Scalar Inference
SOA Stimulus Onset Asynchrony

TCoG Tonal Center of Gravity
TCoG-F Tonal Center of Gravity (Frequency domain)
TCoG-T Tonal Center of Gravity (Time domain)
ToBI Tones and Breaks Indices

GLOSSARY

Buttonbox Specialized hardware used to make button-press behavioral responses. Buttonboxes typically have higher polling rates compared to standard keyboards, allowing for better accuracy in measuring the timing of button presses. The buttonbox used in the present work was a Cedrus RB-740.

Continuum A collection of modified acoustic stimuli consisting of equally spaced manipulations between two endpoint exemplars.

Continuum Step A single contour from a resynthesis continuum, parameterized by one or more manipulated variables (however many is needed to identify a contour uniquely).

Contrast Matrix For a factor with k levels, the contrast matrix is the numeric matrix of size k rows by $k - 1$ columns where each column contains the values needed to encode one planned comparison. The contrast matrix is the generalized inverse of the hypothesis matrix, which is of size $k \times k$. The hypothesis matrix provides information about how each factor level is weighted when computing differences in means (hence informs how regression coefficients are to be interpreted) while the contrast matrix provides the actual numeric values needed to encode these comparisons in a statistical model. See Schad et al. (2020) for additional information.

Edge-tone configuration The present work uses this specifically to refer to the combination of a phrase accent and boundary tone as a unit. For instance, "different edge tones" may refer to the difference between H% and L% or L- and H- (requiring specification of only one level of phrasing) while "different edge-tone configurations" would require the specification of both levels (e.g., L-H%, L-L%, or H-H%).

Excursion In the context of a rising or falling contour, the total difference (in semitones) of the ending F0 value from the accentual F0 value. This is a signed measure: positive values indicate rising F0 excursions and negative values indicate falling F0 excursions. See Chapter 3 for more information.

F0 Sample A single F0 measurement, in Hz, as part of a time-series of multiple F0 samples that comprise a pitch contour.

Helmert Coding A contrast scheme that encodes orthogonal nested comparisons while centering the intercept at the grand mean. For example, if one were to compare the durations of [ʃ,s,n,t] productions, the comparisons would be between (1) [ʃ] and [s] (comparing diffuse vs compact sibilants), (2) {[ʃ],[s]} (the combined mean of the two levels together) and [n] (comparing sibilants to nasals), and (3) {[ʃ],[s],[n]} and [t] (comparing continuants to obstruents). See Sostarics (2024) for more information.

Intonation Following Arvaniti et al. (2022), the linguistically structured and pragmatically meaningful phrase-level modulation of F0. note that the acoustic correlate of intonation is taken to be F0 while the perceptual correlate is pitch.

Nuclear Interval The interval of an intonational phrase from the final pitch accent to the right intonational phrase boundary.

Onglide The region of a pitch contour leading up to a local F0 minimum or maximum corresponding to the pitch accent. For example, H*, L+H*, and L*+H all have rising onglides towards a high accentual target (that differ in shape) while L* has a falling onglide to a low target.

Pitch contour The phonetic expression of a phonologically-specified tune. Measured as a time-series of F0 samples. Used mostly interchangeably with F0 contour, but the latter is used when what is at-issue is the F0 measurements (i.e., Hz values) themselves.

Polling Rate The rate, in Hz, of how many times per second a piece of hardware reports data to the computer. For a standard keyboard the polling rate is usually 125Hz (updates every 8ms) while for a buttonbox the polling rate is somewhere between 320-500Hz (updates every 2-3ms). When the polling rate is lower, then keys/buttons pressed in rapid succession (i.e., faster than the hardware updates) will be reported together and disambiguated by hardware-specific rules. The polling rate affects the precision of RT and RD measurements..

Prosody The collection of suprasegmental features such as intonation, loudness, local tempo, pausing, and voice quality that can be used for linguistic structure or contrast.

Reaction Time (RT) The time, in milliseconds, between the onset of a stimulus and the point at which a participant's behavioral response (e.g., a button press) is made. The precise measurement is affected by the polling rate of the hardware used to make the response (typically either a keyboard or buttonbox). Often analyzed on the natural log scale, where it is referred to as logRT.

Refresh Rate The rate, in Hz, of how many times per second a monitor updates its display. The standard refresh rate is 60Hz, meaning that the image (i.e., the full display of the computer) on the monitor is displayed for $\approx$16.67 milliseconds before it is replaced with a new image from the computer. The monitor used in Chapter 4 has a refresh rate of 165Hz, which allows for lower latency for measurements related to the refresh rate.

Response Duration (RD) The time, in milliseconds, between the point at which a button press is registered and when the button is subsequently released. In other words, the difference in time between a key down event and its corresponding key up event. The precise measurement is affected by the polling rate of the hardware used to make the response (typically either a keyboard or buttonbox).

Resynthesis The process of manipulating a recorded utterance using Pitch Synchronous Overlap and Add (PSOLA) to impose a new pitch contour with researcher-controlled parameters.

RFR-Shaped Tune A nuclear tune with a rising pitch accent and an L-H% edge-tone configuration. This term is intentionally used to capture the fact that multiple phonologically distinct tunes (as predicted by the AM model for MAE) showcase a rise-fall-rise pattern.

Scaled Sum Coding A contrast scheme that encodes deviations between each comparison level and a reference level while centering the intercept on the grand mean. For a factor with k levels, the reference level is coded as $-1/k$ and the comparison level for each contrast is coded as $+(k-1)/k$. Refer to Sostarics (2024) for more information.

Semitones (st) A relative logarithmic value (with a base of 12-root-2) that reflects the interval between two frequencies in a way that is perceptually meaningful and easier to interpret. In music, one semitone is the difference between two adjacent notes (e.g., between B and C or between C and C-sharp). A doubling in frequency is reflected by 12 semitones.

Slope In the context of a rising or falling contour, the slope is the excursion of a rising or falling contour divided by the timespan over which it is expressed. Described in terms of *steepness* or *shallowness*. See Chapter 3 for more information.

Stimulus Onset Asynchrony (SOA) The delay, in milliseconds, between the offset of an auditory stimulus and the onset of a visual stimulus. The precise timing of the SOA is affected by latency from the refresh rate of the monitor presenting the visual stimulus.

Sum Coding A contrast scheme that encodes deviations between each comparison level and the grand mean while centering the intercept on the grand mean. The reference level is coded as -1 and the comparison level for each contrast is coded as $+1$. Refer to Sostarics (2024) for more information.

Trajectory A particular region of interest of a pitch contour showing a monotonic change in F0. For example, the H*H-L% tune has a rising trajectory to a high accentual peak, then a prolonged flat trajectory often described as a plateau.

Tune An abstract phonological expression formulated in terms of a tone sequence of high- and low- tones; specifically, the concatenation of a pitch accent (T*), phrase-accent (T-), and boundary tone (T%) for the nuclear interval of an intonational phrase. Exclusively used in the present work to refer to so-called "Nuclear Tunes," and not larger "sentence-level" tunes that include prenuclear accents.

For David

"The choice of one transcription symbol over another to express an audible phonetic difference inclines us to believe that we are dealing with two different phonological categories. In segmental transcription, this belief is often justified, but in intonational transcription, it actively hinders the development of our understanding."

—*Bob Ladd*

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 1

GENERAL INTRODUCTION

In communication, people convey meaning not only with **what** is said, but also **how** it is said—their INTONATION. In the present work, intonation refers to the linguistically structured modulation of fundamental frequency (F0) in speech; this is to be distinguished from PROSODY, which is taken to further include other suprasegmental features such as duration (i.e., local tempo), voice quality, and intensity. In Germanic languages like Mainstream American English (MAE), German, and Dutch, intonation conveys distinctions in pragmatic meaning such as information structure (Baumann & Riester, 2013; Cole & Chodroff, 2020; Im et al., 2023; Lorenzen et al., 2023; Prince, 1981; Seeliger & Repp, 2023), contrastiveness and focus (Fraundorf et al., 2010; Goodhue, 2022; Krifka, 2008; Repp & Seeliger, 2023; Rooth, 1992; Wagner, 2020; Watson et al., 2008), commitment between interlocutors (Gunlogson, 2001, 2008; Malamud & Stephenson, 2015; Rudin, 2022), intensity and emphasis (Arvaniti et al., 2022; Sandberg, 2024; Watson, 2010), surprisal (Gussenhoven & Rietveld, 2000; Hirschberg & Ward, 1992; Seeliger & Repp, 2023), among many other things (i.a. Büring, 2016; Cole, 2015; Gussenhoven, 2004; Hirschberg, 2017; Ladd, 2008; Pierrehumbert & Hirschberg, 1990; Westera et al., 2021). Unlike distinctions in lexical meaning[1] which form the foundation of contrastive segmental categories in a language, intonation in MAE is somewhat peculiar for its rather variable and flexible range of uses that are highly sensitive to the surrounding discourse context. One context can support or license the use of a variety of intonational patterns yet simultaneously one intonational pattern can also be found in multiple kinds of contexts, suggesting a probabilistic many-to-many mapping between intonational form and function (Roessig et al., 2019; Roettger et al., 2019).

Given such a flexible mapping, it can be difficult to pin down what the contrastive intonational units are in a language. For MAE in particular, the twentieth century saw the development of a

[1] Although, see arguments from Arvaniti et al. (2024) that intonational variation can be modeled in terms similar to the variation present in vowels in the context of Greek pitch accents.

variety of approaches to describe intonational form.[2] A prominent turning point in the analysis of intonation rose from the development of AUTOSEGMENTAL approaches for describing African tonal languages (Goldsmith, 1976, see also Goldsmith, 1990 for additional history). Here, the key insight was that structured patterning of F0-related phenomena (i.e., lexical tones) could be represented on a separate tier from the segmental string while maintaining an association between the two. Pierrehumbert (1980) further extended autosegmental theory to English intonation, providing the basis for AUTOSEGMENTAL-METRICAL (AM) theory (Beckman & Pierrehumbert, 1986; Gussenhoven, 2016; Ladd, 2008), which is now the dominant theory for intonational phonology in the United States and in many places in Europe.

Under AM theory, phrasal intonation is modeled as a sequence of high (H) and low (L) tonal primitives typically annotated with the Tones and Breaks Indices annotation system (ToBI) (Jun, 2022; Silverman et al., 1992; Veilleux et al., 2006). For MAE, the main intonational features are PITCH ACCENTS (H*, L*, or bitonal L+H*, L*+H, H+!H*), which associate with the stressed syllables of words receiving phrasal prominence, and EDGE TONES, which associate with the boundaries of INTERMEDIATE PHRASES (ip, with phrase accents H- and L-) and INTONATIONAL PHRASES (IP, with boundary tones H% and L%), where one intonational phrase is comprised of one or more intermediate phrases. The concatenation of the final, or NUCLEAR, pitch accent in an IP and the edge tones that follow it (e.g., H*L-L%) comprise a nuclear TUNE, which are particularly important in conveying utterance-level pragmatic distinctions (Goodhue, 2024; Pierrehumbert & Hirschberg, 1990; Westera et al., 2021). Tunes with an edge-tone configuration of L-L% have low-falling trajectories that end in a low F0; H-H% yields high-rising trajectories that end in a maximally high F0; L-H% yields trajectories that fall to or maintain a low F0 after the pitch accent, with a rise on the final syllable to a mid-high F0; H-L% is realized in a trajectory that maintains a high plateau following a high pitch accent target, or rises to a mid F0 following a low

[2]Among others, seminal work includes Bolinger (1951), Crystal (1969), Halliday (1967), Jackendoff (1972), and Trager & Smith (1957). Nolan (2022) and Ladd (2015) provide comprehensive historical overviews of the British school of intonational analysis and American level-based systems respectively. The chapters contained within Hirst & Di Cristo (1998) and Gussenhhoven & Chen (2020) additionally provide cross-linguistic descriptions of a variety of intonational systems.

pitch accent target. This work uses the term TUNE to refer to the phonological unit, PITCH CONTOUR to describe its phonetic expression, and TRAJECTORY when narrowly describing a portion of a pitch contour.[3]

A major benefit of the AM model for MAE is that it makes explicit predictions about the intonational contrasts present in the language using a small set of discrete, symbolic, primitives. Based on an inventory of five pitch accents, two phrase accents, and two boundary tones, the AM model thus predicts twenty different phonologically contrastive tunes in MAE. While this set may be smaller than what would be predicted, say, with a comparable system using four levels (see Bolinger, 1951 for a critique of such systems), there is surprisingly limited empirical evidence validating the robustness of the distinctions predicted by the AM model (see also the discussion in Cole et al., 2023, pp. 2–4). In theory, the reduced set of discrete primitives should make it simple to identify and test distinctions among different tunes, but in practice this effort proves to be much more challenging. In addition to the previously mentioned many-to-many mapping between intonational form and function, it has been shown that some distinctions are difficult to perceive or produce (Cole et al., 2023; Dilley & Heffner, 2013; Steffman et al., 2024). Similarly, meaning distinctions between tunes that appear *a priori* clear to the researcher may be revealed to be more mixed when put up to experimental validation (e.g., Buccola & Goodhue, 2023; de Marneffe & Tonhauser, 2019; Watson et al., 2008).

Ideally, a complete theory of a language's intonation would be able to explain which properties of intonational form encode which meaning distinctions—a joint venture between **form** and **function**. Yet research on intonation typically comes from one of two rather isolated perspectives: a "sound" side with a focus on intonational form and a "meaning" side with a focus on intonational function. On the sound side, researchers are primarily concerned with systematic variation in F0: what are the phonologically contrastive categories and how are they cognitively represented, produced, and recovered from the acoustic signal in perception.[4] On the meaning side, researchers are

[3]For example, H*H-L% is a plateau-shaped contour, consisting of a rising trajectory towards a high accentual target (i.e., a rising onglide) followed by a flat and level trajectory until the end of the prosodic phrase.

[4]Relatedly, work on intonation in sociolinguistics frequently investigates questions related to intonational form: How do different varieties of a language differ in their inventory of intonational patterns (e.g., Burdin et al., 2018;

primarily concerned with how, and which, intonational features are used to contribute to categorical distinctions in semantic or pragmatic meaning and, in turn, what these meaning contributions are and how they should be characterized (e.g., are some but perhaps not all distinctions presuppositional in nature or potentially conventionalized) and where intonation fits into existing models of meaning (e.g., the table model of Farkas & Bruce, 2010; see also Malamud & Stephenson, 2015). The assumptions, training, formalisms, methods, rhetoric, venues for publication, and audiences for scholars on either side are very different; work presented to a formal semantics or pragmatics audience is not often concerned with phonetic detail or the phonological specificity of intonation, while work presented to a phonetics or laboratory phonology audience often falls short in identifying an analytic framework for concepts related to information structure and discourse processes. These differences in priorities are not inherently misguided—people should specialize and do what they're trained to do—but it does make it difficult to foster interdisciplinary efforts that actually connect sound and meaning when research is done unilaterally, with only limited involvement of the other side.

As an example of the disconnect between the two lines of work, consider two oft-investigated rising pitch accents in MAE: H* and L+H*. In the ToBI training materials (Veilleux et al., 2006, §2.5.2), the two are differentiated by a "more substantial rising pitch movement" following a preceding low target for L+H*. In the context of intonational meaning and psycholinguistic processing, it is also a common assumption (e.g., Fraundorf et al., 2010; Göbel & Wagner, 2023b; Gotzner et al., 2013) to take H* as the "non-contrastive" accent in contrast to L+H* as the "focus-marking" or "contrastive" pitch accent (as suggested by Pierrehumbert & Hirschberg, 1990, but see critique from Krahmer & Swerts, 2001, pp. 391–393). But whether H* and L+H* comprise one category or two is among the most controversial debates with regard to intonational form (Calhoun, 2004; Ladd, 2022; Ladd & Morton, 1997; Ladd & Schepman, 2003; Orrico et al., 2025; Steffman et al., 2024; Watson, 2010). Empirical work has also repeatedly shown that both H* and L+H* appear probabilistically in the same environments (Bishop et al., 2020; Chodroff & Cole, 2018; Im et al.,

Fletcher & Harrington, 2001; Holliday, 2021) and what might a speaker's use of intonation index socially (Holliday & Villarreal, 2020; Warren, 2016).

2023) and that they overlap in their capacity to convey a contrastive interpretation (Watson, 2010; Watson et al., 2008).[5] Similarly, when annotating recorded materials, it is very well established that adjudicating between H* and L+H* is among the most difficult annotations to make even for expert ToBI annotators (Pitrelli et al., 1994; Silverman et al., 1992; Syrdal & McGory, 2000). Ladd (2022, p. 253), in a critique of ToBI, points out that "the widespread acceptance of ToBI as a standard—together with the fact that the phonetic basis of the distinctions is sometimes readily observable—means that many transcribers of English intonation take it as uncontroversial that there is a categorical distinction between L+H* and H*." Such an assumption of the uncontroversial categorical distinction between the two accents evidently goes beyond just transcription, as exemplified in an investigation of L+H* with regard to mirativity by Rett & Sturman (2020, p. 17), where L+H* "has the additional virtue of being categorical; [...] a construction either has or does not have an L+H* pitch accent." The disconnect here is that the uncertainty in the investigation on intonational form (i.e., with respect to L+H* and H*, is it one category or two?) is not always taken under consideration in investigations of intonational meaning.

1.1 Goals and Overview

Despite the AM model's widespread adoption, evidence for whether the full predicted inventory of tunes are robustly different from one another in perception, production, and interpretation is highly variable. The goal of this thesis is to assess the categorical status of a subset of tunes predicted by the AM model for MAE by shining a light on two areas of intonational meaning research that stand to benefit from reconnecting with rigorous investigation of intonational form. The focus here is in relation to between- versus within-category variation in intonational form. While this thesis and the questions that guide it are thus very much on the "sound side," distinctions in form are investigated on the basis of their interpretation and processing. This thesis takes a variety of approaches across thirteen perception/comprehension experiments, where the choice of method is informed by the kinds of meanings different tunes have been described to convey. The structure

[5] Similar results have also been found in German (Baumann & Riester, 2013; Roessig, 2021; Roessig et al., 2019; Seeliger & Repp, 2023)

of this thesis can be described in terms of two parts, with Part 1 investigating variation in RISING DECLARATIVES and Part 2 investigating variation in the RISE-FALL-RISE tune.

Part 1 of this thesis, comprised of Chapters 2 and 3, looks at RISING DECLARATIVES, or, the use of rising intonation with declarative sentences. Broadly speaking, rising declaratives such as *Molly's from Branning?* are used to ask a question about something that someone has some contextual evidence for. While the canonical intonational tune for such utterances is taken to be L*H-H% (Bartels, 1997; Gunlogson, 2001; Pierrehumbert & Hirschberg, 1990; Rudin, 2022, among many others), more recent work has argued that H*H-H% serves to convey an assertive type of rising declaratives (Jeong, 2018) in contrast to the inquisitive L*H-H% (see also Hirschberg & Ward, 1995 who offer a slightly different analysis of H*H-H%). While the distinction between rising and falling intonation (i.e., H*L-L% versus L*H-H%) is uncontroversial, it is more difficult to find evidence that L*H-H% and H*H-H% are actually interpreted differently from one another in MAE (c.f. targeted investigation of analogous tunes in Dutch from Gussenhoven & Rietveld, 2000) despite apparent robustness in perceptual discrimination (Cole et al., 2023). An alternative proposal is that the distinction between the two may not be linked to a category-level distinction on the basis of the pitch accent specification, but instead linked to (potentially paralinguistic) variation within a single category (e.g., in the scaling of the final F0, Goodhue, 2024), of which L*H-H% and H*H-H% may be extremes along a gradient continuum (Ladd, 2022, p. 243). Chapter 2 presents a series of perception experiments that seek to assess whether variation in the inquisitive/assertive interpretation is more closely linked to a between-category distinction (i.e., indicating between-category variation of H*H-H% versus L*H-H% vis-à-vis falls like H*L-L%) or whether variation in the ending F0 is responsible for this contrast (indicating within-category variation related to H%). Chapter 3 presents a reanalysis of the experimental data in Chapter 2, offering a comparison of different models that use three different acoustic measures related to rises and falls: F0 excursion, slope, and Tonal Center of Gravity (Barnes et al., 2021).

Part 2 of this thesis, comprised of Chapter 4, looks at the RISE-FALL-RISE tune (henceforth RFR). Unlike rising declaratives, which can be conveniently given the non-technical elevator pitch

of "intonation used to ask a question with declarative syntax," RFR is a bit more difficult to describe in part due to the multitude of competing accounts that describe what its meaning contribution is. Under some accounts, RFR conveys that the speaker is uncertain and does not wish to commit himself to some salient scalar alternative (Ward & Hirschberg, 1985) and so serves as a bit of a "polite hedge" (Ronai & Göbel, 2024, p. 8). Under other accounts, RFR conveys that there exist alternatives that remain disputable (Constant, 2012) or should be rendered salient (Göbel, 2019), or that there exist alternative questions or speech acts (Büring, 2003; Wagner et al., 2013; Westera, 2019). Some work goes as far as to equate RFR with the marking of contrastive topic (Büring, 2003; Constant, 2014), while others are adamant that RFR and contrastive topic are orthogonal to one another (Wagner, 2012). Two things stand out in this literature: A persistent connection to alternatives and a focus on *the* RFR tune. Yet, what is offered as an annotation of RFR under AM terms is inconsistent across the literature, with researchers varying in whether L*+HL-H% is entirely separate from (L+)H*L-H% or whether the differences are merely paralinguistic and reflect within-category variation of a broad RFR class. Chapter 4 presents a series of perception experiments that investigate the potential differences between these "RFR-shaped tunes," which putatively differ in their pitch accent specification, in the context of scalar inference (SI, Horn, 1972) in both offline interpretation and online processing.

Although both rising declaratives and RFR have long histories of prior work, they are fairly disjoint from one another. The link between the two, for the purposes of the present work, is that the AM model predicts categorical distinctions between different types of rises (e.g., L*H-H% and H*H-H%) and between different types of RFR-shaped tunes (H*, L+H*, and L*+H as they combine with L-H%). When looking at these different tunes in the context of how their meaning contributions have been previously described, whether these category-level distinctions matter is unclear. As previously mentioned, the goal of this thesis is not to propose new characterizations of the pragmatic contributions of rising declaratives nor RFR-shaped tunes, but rather to take insights from the pragmatics literature on these phenomena (e.g., speaker inquisitiveness for rises and scales for RFR) to then more rigorously look at intonational form. An extensive literature review of the

two phenomena under investigation is relegated to Chapter 2 (for rising declaratives) and Chapter 4 (for RFR). Chapter 5 concludes the thesis and relates the main findings of both parts to the question of between- and within-category variation.

Chapter 2

VARIATION IN RISING AND FALLING INTONATION

2.1 Introduction

It is generally well known that intonation in Mainstream American English (MAE) conveys distinctions in pragmatic meaning (Bartels, 1997; Büring, 2016; Cole, 2015; Hirschberg, 2017; Pierrehumbert & Hirschberg, 1990; Prieto, 2015; Westera, 2017; Westera et al., 2021). Yet, at the same time, intonation has also been shown to display significant variation in both its form and the interpretation of its many functions, displaying a many-to-many mapping on top of overlapping or "fuzzy" boundaries between phonologically contrastive categories (Arvaniti, 2019; Arvaniti et al., 2022; Ladd & Schepman, 2003; Roettger et al., 2019; Watson et al., 2008). In perception studies, participants are often able to accommodate similar interpretations for intonational contours that are hypothesized to be distinct in their meaning contributions (Buccola & Goodhue, 2023; Cole, 2015; Nilsenová, 2006) though predicted associations may nonetheless be revealed in probabilistic tendencies (Seeliger & Repp, 2023; Sostarics et al., 2025). In production studies as well, participants often produce a variety of intonational contours in contexts that *a priori* should license the use of only specific contours (Goodhue et al., 2016; Ronai & Göbel, 2024, see also Chodroff & Cole, 2018, 2019 for variation in the production of pitch accents specifically). Even within the same speaker, there can be notable variation in the phonetic expression of a single target intonational tune yet simultaneously only small differences between two tunes that are proposed to be categorically distinct in their phonological specification (Steffman et al., 2024). This variation at both the phonetic and pragmatic/interpretational levels is in stark contrast to the segmental domain, where the contrast between two phonemes can be established on the basis of a distinction in lexical meaning. For example, having two distinct lexemes *bear* and *pear* entails a categorical difference between phonemes /b/ and /p/. Identifying the analogous intonational features that minimally contrast along particular meaning dimensions is much less straightforward (though cf. arguments from Arvaniti et al., 2024 with reference to Greek).

Where should one start to look for contrastive intonational features? One oft-discussed distinction in English is the interpretational contrast with declarative sentences when uttered with different intonational contours. Specifically, a speaker uttering a declarative sentence with falling intonation (e.g,. *It's raining.*) is taken to express an ASSERTION while a speaker uttering the same sentence with rising intonation (e.g,. *It's raining?*) is typically taken to express a QUESTION (Bartels, 1997; Farkas & Bruce, 2010; Goodhue, 2024; Gunlogson, 2001, 2008; Jeong, 2018; Malamud & Stephenson, 2015; Rudin, 2022).[1] In MAE, the combination of declarative sentence type and rising intonation is typically referred to as a RISING DECLARATIVE (RD). The present work refers to the contrast between these questioning and asserting interpretations as the Q/A CONTRAST.

Recent work on rising declaratives has suggested that the distinction between monotonically rising pitch versus monotonically falling pitch is too coarse of a parameterization to account for differences in interpretation. In a study of intonational meaning, Jeong (2018) finds that declaratives with shallower (i.e., less steep) rises are probabilistically more likely to receive assertive interpretations rather than INQUISITIVE (i.e., question) interpretations.[2] In contrast, steeper rises were more likely to receive INQUISITIVE interpretations. Thus Jeong proposes a distinction between shallow ASSERTIVE rising declaratives (ARDS) and steep INQUISITIVE rising declaratives (IRDS).

In related work building on the account proposed by Jeong (2018), Goodhue (2024) has argued for further distinctions within the IRD class. CONFIRMATIVE rising declaratives (CONFRDS) are used to confirm something the speaker is already relatively certain about, but notably the speaker is still asking their addressee to confirm, hence CONFRDS are INQUISITIVE in nature. These CONFRDS are described as being shallower in slope, like ARDS, hence accounting for the at-

[1] It is not uncommon to see rising declaratives referred to as "polar question intonation/rises" (e.g,. Goodhue, 2024), but it should be noted that rising declaratives (a sentence type plus a rising tune) and polar interrogatives (a sentence type, which can occur with rising or falling intonation) are not equivalent (Geluykens, 1988). However, much work takes the denotation of both RDs and polar interrogatives to be $\{p, \neg p\}$ with the intuition being that RDs convey a kind of biased question while polar interrogatives are more neutral (see Rudin, 2022, pp. 343–344 for additional review and a series of empirical generalizations related to the distribution and use of RDs, building off of observations from Farkas & Roelofsen, 2017).

[2] Note that in both Jeong (2018) and the present work, "assertive interpretation" refers to the interpretation that the speaker is making an assertion and not that the speaker is being forceful. The use of "inquisitive" here is also adopted from Jeong's work to refer to questioning interpretations.

chance levels in ASSERTIVE versus INQUISITIVE interpretations in Jeong's experiment—there are competing INQUISITIVE and ASSERTIVE interpretations of the shallow rise. The IRDs with the steepest F0 rises are linked to INCREDULOUS rising declaratives (INCRDS), where the speaker is not merely questioning something but is expressing incredulity, surprise, or disbelief about some information.

Thus far this taxonomy of rising intonation is described in terms of variation in steepness, but there are multiple ways in which steepness can vary as a result of phonological distinctions and phonetic variation. Indeed, Goodhue (2024, p. 8) provides three potential sketches for describing the distinctions between these subtypes of RDs, largely boiling down to a question of whether (and which) RD subtypes are phonologically distinct in their tonal specification or whether paralinguistic factors drive differences in the phonetic expressions (in particular for INCRDS). The latter paralinguistic proposal is motivated based on prior work showing that more extreme pitch excursions convey greater speaker arousal, charisma, engagement, and emotional activation (Gussenhoven, 2004; Ladd et al., 1985; Niebuhr et al., 2018). Rather than taking steepness as the primary cue to interpretation, this work takes steepness as an epiphenomenal stepping stone with the goal of considering a broader range of acoustic dimensions—which, in turn, affect steepness—that relate to different phonological distinctions in intonation.

2.1.1 Zooming Out from Steepness

This section sketches a series of "fundamental parameters" by which rises (and, by extension, falls) can vary, making reference to particular dialects of English where these parameters have been investigated. Specifically discusssed here is variation in where rises start and end with reference to both fundamental frequency (F0) value and time, where the latter motivates an investigation of both temporal extent and the alignment of F0 targets. This work will make reference to how variation in different parameters impacts the steepness of a rise for expository purposes, but it is important to keep in mind that steepness is being used as a guide towards identifying potential

parameters that may be meaningful in the context of rising and falling intonation.[3]

First, rises can vary in the F0 of the starting point of the rise while rising to the same final F0 target, where lower starting points will lead to steeper rises. Variation in the starting F0 target of the rise has been found to be important for the Q/A contrast in Australian English (Fletcher & Harrington, 2001). In MAE, Cole et al. (2023) show evidence from an imitation paradigm that naïve speakers are able to reproduce distinctions between rises that phonologically differ in this dimension (though c.f. more mixed results in Steffman et al., 2024). Similarly, Jeong (2018) proposes a phonological distinction, differentiating ARDs from IRDs, along this dimension. In Dutch, a closely related West Germanic language with an intonational system similar to English, there is also perceptual evidence for two rises that vary along this dimension (Gussenhoven & Rietveld, 2000).

Second, rises can vary in the F0 of the ending point of the rise while rising from the same starting F0 target, where higher ending points will lead to steeper rises. As previously discussed, variation in the ending F0 target has been described to be important for the Q/A contrast in MAE. This parameter has also been found to vary systematically for New Zealand English speakers (Warren & Fletcher, 2016) for the Q/A contrast and has been shown to relate to positive/negative bias in the expressed polar question in Canadian English (Arnhold et al., 2021).

Third, if the starting and ending points are both held constant, but the duration over which the rising excursion takes place is adjusted, then this duration manipulation will also affect the steepness of the rise. For example, a $+6$ semitone rise over 100ms will be more steep than the same magnitude rise over 200ms. There is a related question of whether and how listeners normalize speech rate (and by extension, how normalization is done for different speakers), but this question is set aside in this work (though see Kurumada & Buxó-Lugo, 2024; Xie et al., 2021).

Fourth, the rise can vary in the alignment of the starting point such that later-aligned starting points will lead to steeper rises. In much the same way as the previous example, if a rise begins

[3]It has been reported that some varieties of English, notably African American English, do not use rising intonation with declarative sentences to convey polar questions (Conner, 2020). The scope of investigation here is on how variation in rising and falling F0 contours affects the interpretation of declarative sentences, and questions about the typology of linguistic encodings for polar questions/biased questions are left for future work.

earlier, then it will have a longer duration over which the rise takes place; if the rise begins later, then it has comparably less duration and thus will be more steep. Alignment has been found to be important for New Zealand English speakers (Warren, 2014; Warren & Fletcher, 2016).

Finally, while the focus thus far has been implicitly on rising and falling trajectories arising via straight-line interpolation between F0 targets, one might also ask whether the shape of the trajectory from one point to another is at all important. For instance, the curvature of different F0 trajectories in MAE has been shown to affect judgments of perceptual similarity with corresponding systematic differences in production (Barnes et al., 2021; Steffman et al., 2024).

When considering these parameters, we can note two properties. First, they can, in principle, vary independently of one another.[4] Second, these are the same parameters by which falling trajectories can vary: a fall can start or end higher or lower, earlier or later. So, when discussing variation in "steepness," what are the relevant parameters that are actually being varied? Evidently, there are multiple moving parts under the hood that may need to be considered in a successful parameterization of shallow versus steep rises and falls, but how do these parameters relate to phonological structure?

Autosegmental-Metrical (AM) theory (Beckman & Pierrehumbert, 1986; Ladd, 2008; Pierrehumbert, 1980) offers one such phonological parameterization of these dimensions, where rises and falls (among other phrase-final pitch patterns) are reduced to a limited inventory of high and low tonal targets. Recall from the general introduction (Chapter 1) that under the AM model for MAE, pitch accents (H*, L*, L+H*, L*+H) associate with the stressed syllable of a word while edge tones associate with the right edges of prosodic phrases; the concatenation of the final pitch accent in an intonational phrase and the edge tone configuration that follows it comprise the nuclear tune. This work will focus on four pitch accents—monotonal H* and L* and bitonal L+H* and L*+H—as they combine with (primarily) H-H% and L-L% edge-tone configurations,[5] yielding a variety of rising and falling contours. The link between these phonological constructs and the

[4]Though see Iskarous et al. (2024) for discussion that peak alignment, peak height and rise curvature are linked.

[5]Recall that tunes with an edge-tone configuration of L-L% have low-falling trajectories that end in a low F0; H-H% yields high-rising trajectories that end in a maximally high F0; L-H% yields trajectories that fall to or maintain a low F0 after the pitch accent, with a rise on the final syllable to a mid-high F0; H-L% is realized in a trajectory that

fundamental parameters previously described are discussed further below.

The starting point of a rise or fall systematically varies with the phonological choice of pitch accent as well as with the phonetic scaling[6] of that pitch accent. For instance, a rise may start from a high point (H*) or a low point (L*), which may be variably scaled progressively higher or lower. Variation in the alignment of this starting point is also implicated in the distinction between L+H* and L*+H, where the accentual peak for L+H* is realized earlier than that of L*+H. A rise may also end at a high F0 value or a lower F0 value, though whether this is best described as a phonological distinction (e.g,. between H-H% and L-H%) or as phonetic variation within H-H% is an open question (see also Goodhue, 2024).

Within an AM parameterization, and with regard to the steepness of a rising contour, Jeong (2018, p. 312) hypothesized that "the slope of the rise, which is often determined by the relative position of the nuclear pitch accent, is the most relevant indicator of the ASSERTIVE vs. INQUISITIVE rising declarative distinction," thus linking L*H-H% to IRDs and H*H-H% to ARDs. However, the materials used in that study did not manipulate the starting F0 target of the nuclear pitch contour, but rather the final F0 target (i.e., the ending point varied but the starting point did not), and so Jeong's results are potentially better understood in terms of variation in the scaling of the boundary tone.[7] Regardless, the level of analysis that Jeong (2018) takes is at the level of the holistic tune, not the tones that comprise them, and so it suffices to say that the meaning distinction described in Jeong (2018) distinguishes between two distinct rising patterns. This work does not

maintains a high plateau following a high pitch accent target, or rises to a mid F0 following a low pitch accent target.

[6]This work uses *scaling* to refer to "vertical" variation in F0 of a high or low tonal target. For instance, H* and L* always differ in that H* is higher than L*, but H* and L* may be variably higher or lower respectively. Relatedly, L* and low edge tones have been observed to be more limited in their capacity to lower (e.g., under pitch range expansion), see discussion in Liberman & Pierrehumbert (1984), Pierrehumbert (1980), and Seeliger & Repp (2023) and also Gussenhoven & Rietveld (2000) for related variation in Dutch.

[7]Jeong (2018, p. 316) argues that this manipulation of the materials may be alternatively understood in terms of reanalysis of the speaker's pitch range on the basis of the final rise (see also discussion in Goodhue, 2024, p. 6 footnote 7). For instance, a speaker with a steep final rise may have the "mid-level" prenuclear F0 reanalyzed as being low pitch in order to accommodate the large rising pitch excursion, leading to a phonological analysis of the nuclear pitch contour as L*H-H%. A shallower rise would not require an analysis of such a low starting point, and so the mid-level pitch can be analyzed as just that, leading to a phonological analysis of (!)H*H-H%. While this helps to salvage the choice of L*H-H% versus H*H-H% labels, allowing for a comparison with prior work on H*H-H% like Hirschberg & Ward, 1995, the same manipulation can also be understood in terms of L*H-H% versus L*L-H% (as the materials are restricted to 2-syllable words), allowing for a tune-level distinction to be maintained.

dispute that two rising intonational patterns can have relatively stable semantic/pragmatic contributions nor does it propose a refinement to the formalism Jeong provides. Rather, it remains an open question whether the variation in the region of the contour corresponding to the pitch accent plays a robust role in interpretation along the ASSERTIVE/INQUISITIVE dimension, or whether the distinction Jeong describes is better understood in other terms.

There is a broader question here regarding between- and within-category variation, in particular as it relates to the perceptual reality of putatively different rises. Although L*H-H% and H*H-H% are predicted to be contrastive units by the AM model for MAE and have been described as having rather different interpretations (Hirschberg & Ward, 1995; Jeong, 2018), in truth there is little work rigorously investigating such predicted distinctions in intonational form. Cole et al. (2023) reports a clear distinction between H*H-H% and L*H-H% in imitated productions, but Steffman et al. (2024) report that imitations of H*H-H% and L*+HH-H% are very similar to one another and are discriminated below chance in perception.[8] Similarly, Dilley & Heffner (2013) show that imitations of rising contours that vary in the alignment of the low valley (from early to late) show lower accuracy in contrast to imitations of falling contours that vary in their peak alignment. While Warren & Fletcher (2016, pp. 33–45) reviews a number of studies indicating that the starting point of a rise can vary (particularly across dialects of English), Ladd (2022, p. 253) argues that the widespread practice of adopting ToBI labels to (conveniently) "represent audible differences that convey an intonational nuance" (i.e., L*H-H% versus H*H-H%) may inadvertently entail phonological distinctions across categories when the observed phonetic variation may "actually involve the extremes of a gradient continuum." In other words, while there may be slight nuances to the interpretation of various rising contours, whether these map 1-to-1 to phonologically contrastive tunes ({L*/H*/L+H*/L*+H}H-H%) or is reflective of meaningfully gradient phonetic variation under a broader rising category remains unclear. Cole et al. (2023) additionally raise the possibility that perhaps the pitch accent distinctions are not robust in the context of a tune using an H-H% edge-tone configuration; in appealing to the segmental domain, it is possible that the phonological

[8]Note however that these two studies implement the H*H-H% tune in their materials in slightly different ways.

contrast between pitch accents is neutralized in an H-H% context.

The goal of this chapter is to take a closer look at variation in the phonetic expression of rises and falls by manipulating various "fundamental parameters" by which rises and falls can vary (starting/ending F0, alignment, shape) as they relate to proposed phonological (i.e., category-level) distinctions predicted by the AM model for MAE. In broad terms, the research question is thus: what region of a rise or fall over the nuclear interval (from the nuclear pitch accent to the right edge of the intonational phrase) matters for the contrast between ASSERTIVE and INQUISITIVE interpretation in the context of declarative syntax? Is it only the final F0 targets corresponding to the edge tones that matter, or does the region corresponding to the pitch accent also matter for this meaning distinction? Prior work probing the contrast between falling and rising intonation has focused on a limited phonetic space (Jeong, 2018) or has manipulated both the pitch accent and boundary tone simultaneously (see Xie et al., 2021, p. 13 Fig. 12), making it difficult to identify the separate contributions of the pitch accent and edge tones. In contrast, the present work tests meaning distinctions for a variety of rising and falling F0 contours that reflect variation not only **across** putative categories (e.g., H*H-H% and L*H-H%) but also **within** these categories through gradient phonetic continua.

This chapter presents results from seven perception experiments using a two-alternative forced choice (2AFC) task. Each experiment makes use of a different 25-step crossed continuum created by manipulating parameters related to the pitch accent (scaling, alignment) separately from boundary tone. Experiments 1 through 2c investigate contours using monotonal pitch accents (L* and H*). Experiment 3 investigates scaling in the accentual peak of contours using the bitonal L+H* pitch accent. Experiment 4 investigates differences in alignment between two bitonal pitch accents—an early-aligned L+H* and a late-aligned L*+H. Experiment 5 extends the 2AFC paradigm to elicit free-text responses for an exploratory look at the range of interpretations that are salient to participants. The results show that variation in the ending F0 target (i.e., boundary tone choice and scaling) is the most robust cue for the Q/A contrast. Variation in the region associated with the pitch accent plays a more limited role, often introducing orthogonal dimensions of

meaning (such as focus) that can interfere (in the experimental task presented here) with narrowly judging whether the speaker is asserting or questioning.

2.2 Materials Overview

The materials in this work rely on fine-grained control of the acoustic targets used in various rises and falls. This section will go into detail regarding the overall process of creating the materials for this chapter, with particular attention to the recording process, the duration manipulation used to standardize the durations of the stimuli, and a guiding framework for how the pitch manipulations will be discussed moving forward. Rather than describe the F0 manipulations for all seven experiments together, subsequent sections for each experiment will describe the F0 manipulations for their respective stimuli. Thus, this section serves to describe what is **common** about the materials across all experiments.

The materials used here can be seen as extending those of Jeong (2018). The stimuli are comprised of five declarative sentences ending in two-syllable words of the form *Name's* {*determiner / preposition*} *noun*, for example, *Molly's from Branning*.[9] The final noun always had word-initial stress and is comprised entirely of voiced segments to facilitate pitch measurement and resynthesis.

2.2.1 Control Over F0 and Duration

The goal was to create standardized rising and falling contour shapes that are acoustically comparable across the sentences. As previously discussed, this involves manipulation of the value of the starting/ending F0 targets, their alignment, and the shape of the trajectories between the two points. These manipulations are implemented using pitch resynthesisis (PSOLA) while standardizing the duration of the nuclear accented words in Praat (Boersma & Weenink, 2020b).[10]

[9]The other sentences are *Gavin's on broadway; Megan's a grandma; Ryan's in Greenview;* and *Joey's from Bronville.*

[10]Recall that the slope of a rise can vary by changing the time over which the rising pitch excursion takes place. Accordingly, duration needs to be controlled across the sentences to prevent idiosyncratic differences in local tempo from confounding the F0 manipulations.

2.2.2 Approach to Resynthesis

The five sentences were recorded multiple times with a variety of target intonational tunes[11] in a soundproof recording booth using a Shure SM27 microphone. Approximately 300-400 tokens were originally recorded and then manually inspected to determine suitability for pitch resynthesis, yielding a final candidate corpus of 223 tokens. The raw pitch contours for this corpus are shown in Figure 2.1, where the overall pitch range is approximately 60-180Hz.

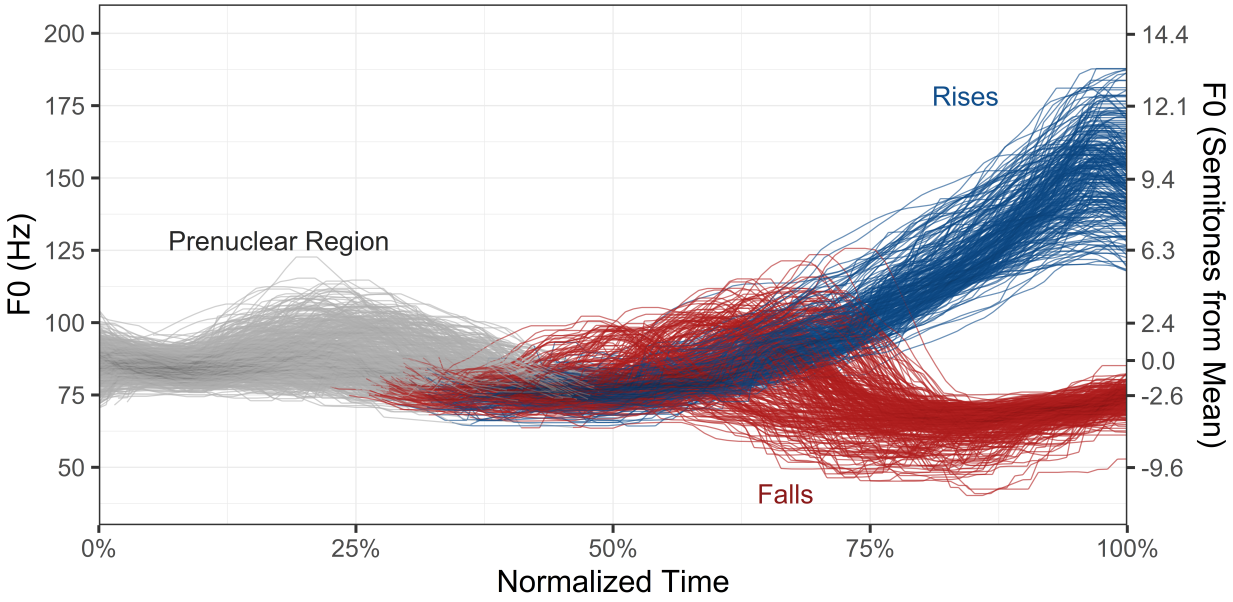


Figure 2.1: Raw F0 contours for the corpus of falls and rises.

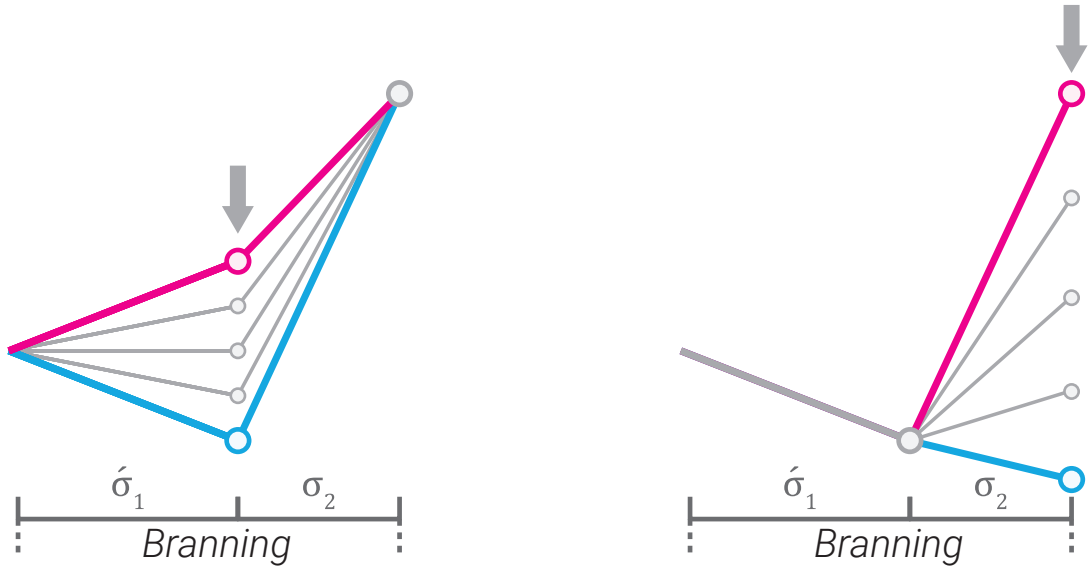
These recordings were then force-aligned using the Montreal Forced Aligner (MFA) (McAuliffe et al., 2017a) to obtain duration estimates of each syllable of the nuclear-accented word. The first syllable had an average duration of 311ms and the second syllable had an average duration of 364ms. The force-aligned textgrids for the recordings with the closest durations to these average values were then manually checked and corrected. These corrected files were then segmented into

[11] Using ToBI labels, the sentences were recorded with H*L-L% (falling), L*H-H% (rising), L+H*L-L% (falling with an early-aligned and domed peak), and L*+HL-L% tunes (falling with a later-aligned and scooped peak). An additional set of materials with three-syllable words was also constructed but ultimately not used to maintain comparability with Jeong's materials. The goal with amassing recordings from a variety of target intonational tunes was to get an idea of an appropriate range for the manipulations and also to ensure that the source recordings yield natural-sounding results even when resynthesized to a drastically different contour.

syllables and manipulated to be equal to the calculated averages.[12] The final files were then manually inspected by myself for resynthesis quality and naturalness; ultimately, recordings uttered with H*L-L% intonation were selected for further F0 manipulation. In summary, there are five declarative sentences that have been duration manipulated such that the syllable lengths of the nuclear accented word are the same for every sentence, which ensures that the resynthesized pitch contour continua for one sentence are exactly the same as those for every other sentence.

Following the above-described duration manipulation, the materials were then resynthesized to new F0 contours. Recall that we can parameterize falling and rising pitch contours minimally with two parts: the point at which the rise/fall starts and the point at which it ends. Phonologically, these can be described in terms of the pitch accent and edge tone(s) respectively. Moving forward, we will refer to the "starting point" as ACCENTUAL PITCH: the F0 value taken to be the acoustic target of the pitch accent. Generally, the alignment of high accentual peaks are closer to the end of the stressed syllable, not the start; accordingly, the accentual pitch target is aligned to the **end** of the nuclear-accented syllable—approximately halfway through the phrase-final word. This means that for a high accentual pitch target, there will be a rising onglide to the accentual pitch target (and conversely a falling onglide for low accentual pitch targets). The "ending point" of the fall/rise will similarly be referred to as ENDING PITCH: the F0 value taken to be the acoustic target of the edge-tone configuration, which is straightforwardly aligned with the end of the nuclear word. These two points can be manipulated separately, but **together** determine the magnitude of the pitch excursion (the difference in F0 from the ACCENTUAL PITCH target to the ENDING PITCH target) and the slope of the contour. **Falling** contours thus refer to contours where F0 decreases from the accentual pitch target to the ending pitch target (i.e., a negative pitch excursion) while **rising** contours thus refer to contours where F0 increases from the accentual pitch target to the ending pitch target (i.e., a positive pitch excursion). Figure 2.2 depicts schematic versions of the separate of accentual pitch and ending pitch continua across the nuclear interval of the prosodic phrase (i.e., over *Branning*).

[12]The resynthesized falls with linear trajectories from the accentual peak to the end of the word did not sound characteristic of typical H*L-L% intonation when using the full 364ms long syllable. To avoid such unnaturalness while keeping the linear trajectory, the second syllable was shortened to 70% of the average duration, which sounded more natural while retaining the linear trajectory.



(a) Schematic accentual pitch targets

(b) Schematic ending pitch targets

Figure 2.2: (a) Schematic manipulation of accentual pitch, holding the ending pitch target constant. (b) Schematic manipulation of ending pitch, holding the accentual pitch target constant. Highlighted are the high (pink) and low (blue) endpoints of each continuum.

Each experiment specifies two separate continua: a 5-step continuum for accentual pitch and a 5-step continuum for ending pitch, which are then fully crossed. Thus, each experiment will have 25 **phonetically** distinct pitch contours, but the number of **linguistically contrastive** tunes these pitch contours map onto is (likely) fewer, with at minimum a contrast between rising and falling contours (Cole & Steffman, 2021; Cole et al., 2023).

2.3 Hypotheses

The primary research question in this chapter is: which intonational features (as expressed through the fundamental acoustic parameters previously described) are relevant to the Q/A contrast for MAE in the context of falling/rising declaratives. From a holistic viewpoint, treating the relevant unit of pragmatic analysis as the whole tune, this becomes a question of which distinctions are conveyed through phonologically different tunes (e.g., the distinction between H*H-H% and L*H-H%) and which distinctions reflect meaningful within-category variation (e.g., scaling in the phonetic expression of L*H-H%). From a compositional viewpoint, where the relevant unit of

analysis is the individual intonational features (e.g., pitch accent and boundary tone) that comprise tunes, this becomes a question whether the relevant distinction is restricted to the edge tones or whether the pitch accent additionally plays a role beyond the orthogonal referential meaning distinctions they are taken to convey (e.g., the seminal account of Pierrehumbert & Hirschberg, 1990, described further below). These need not necessarily be mutually exclusive, as the (potentially conventionalized) range of inferences licensed by the use of a tune may or may not go beyond what would be predicted by compositional descriptions of individual tonal features. Two hypotheses are proposed:

H1. **Edge-only Hypothesis**: The Q/A contrast is conveyed solely by the edge-tone configuration of the nuclear tune, which is cued by the F0 trajectory following the accentual peak towards the ending F0 target. Interpretation is associated solely with variation in the choice (i.e., L-L% or H-H%) and scaling of the ending pitch target.

H2. **Integrative Hypothesis**: The Q/A contrast is conveyed jointly by the pitch accent **and** the edge-tone configuration.

If the pitch accent were to matter for interpretation of the Q/A contrast, what effect might we expect it to have? Seminal work taking a compositional approach to tune meaning primarily derives the difference between H* and L* based on the contrast between L*H-H% and H*L-L%, where both the pitch accent and the edge tones differ (Pierrehumbert & Hirschberg, 1990). Broadly, H* is described as conveying new information while L* is used with given information or information that is salient but not part of the current predication (Hobbs, 1990; Truckenbrodt, 2012; Westera et al., 2021). If we take L*H-H% and H*L-L% as the canonical intonational forms for denoting INQUISITIVE rising declaratives and non-INQUISITIVE falling declaratives, respectively, then we might expect trajectories that deviate from canonical expressions of L*H-H% to sound less INQUISITIVE, while trajectories that deviate from canonical expressions of H*L-L% would sound less ASSERTIVE.[13] This account predicts H*H-H% to sound more ASSERTIVE than L*H-H%,

[13]Here, saying that a tune "sounds more or less ASSERTIVE/INQUISITIVE" with a particular phonetic expression should be taken to mean that that phonetic expression is **less likely to yield an ASSERTIVE/INQUISITIVE interpretation**, and not as a claim related to potential degrees of assertiveness (Wolf, 2014) or inquisitiveness (beyond what's described by different types of questions in Groenendijk & Roelofsen, 2009) nor as a claim related to degrees of commitment (Mazzarella et al., 2018).

which would be in line with Jeong (2018) and with description from Hirschberg & Ward (1995) arguing that the H* accent marks new information to be added to the interlocutor's beliefs. However, this account also makes the odd prediction that L*L-L% would sound more INQUISITIVE than H*L-L%, which intuitively does not appear to be true (e.g., in the discourse contexts describe in Sostarics & Cole, 2021). A more plausible prediction is that there is little to no variation in the interpretation of falling contours related to the choice of H* vs. L* pitch accent, but there may be variation for the rises: higher accentual pitch may lead to a greater likelihood of ASSERTIVE, rather than INQUISITIVE, interpretations.

2.3.0.1 The Role of Bitonal Pitch Accents

While we might expect an increasing likelihood of ASSERTIVE interpretations when investigating variation between L* and H*, it is not immediately clear how the bitonal accents, L+H* and L*+H, would pattern. The L+H* pitch accent is frequently linked to the prosodic marking of focus (Rooth, 1992; Wagner, 2020), which can be described as indicating "the presence of alternatives that are relevant for the interpretation of linguistic expressions" (Krifka, 2008, p. 247). The L*+H accent has been described as invoking a scale (Pierrehumbert & Hirschberg, 1990) or imposing an ordering relation on the evoked set of focus alternatives (Göbel, 2019). When limiting the scope to contrastive focus, prior work has shown that L+H* and H* overlap in their capacity to convey a contrastive interpretation (Watson et al., 2008), suggesting a link to prominence more broadly (Watson, 2010).

In terms of prominence distinctions, multiple studies have found a probabilistic association between pitch accents and information structure in both German (Baumann & Riester, 2013; Repp & Seeliger, 2020; Roessig, 2024; Roessig et al., 2019) and MAE (Chodroff & Cole, 2018, 2019; Im et al., 2023; Roettger et al., 2019). Im et al. (2023) further propose a continuous and probabilistic relationship between an Accentual Prominence hiearchy on the one hand (where L* $<$ H* $<$ L+H* $<$ L*+H) and a Givenness hierarchy on the other (where given $<$ bridging $<$ unused $<$ new). Such a hierarchy is also reflected in alternative approaches to representing pitch accent categories in

terms of continuous variation, such as in dynamical systems approaches where the different pitch accents arise from principled variation in a single parameter of the system (Iskarous et al., 2024).

To summarize, the bitonal accents have been described as relating to scales and contrastive focus marking, but work on prominence has repeatedly shown that a simplistic and deterministic mapping between, say, L+H* and contrast is not supported. Based on the Accentual Prominence hierarchy, we might hypothesize that the likelihood of ASSERTIVE interpretations increases from L* to H* and may extrapolate to further increased likelihood of ASSERTIVE interpretations for L+H* and L*+H. Alternatively, it may be that as prominence of the pitch accent increases, the likelihood of an interpretation enriched by the marking of focus may also increase. However, it is an empirical question of whether such enrichment would serve to **enhance** the Q/A contrast, or whether it would detract (bringing response proportions closer to chance) instead. This question will be specifically addressed in Experiment 3.

2.4 Monotonal Pitch Accents (Exp. 1-2c)

The first experiment investigates variation between the monotonal high and low pitch accents (H*/L*) in rising and falling tunes. The goal of this experiment is to relate variation in the interpretation of an utterance as ASSERTIVE or INQUISITIVE with variation in the pitch contour over the nuclear interval (i.e., the nuclear tune), as parameterized by accentual pitch (as the cue to the pitch accent) and ending pitch (as the cue to the edge-tone configuration).

2.4.1 Materials

Three simplifying constraints based on the materials used in Jeong (2018) are adopted when creating the materials.[14] First, the prenuclear region of the sentence (e.g., *Molly's from* in *Molly's from Branning*) is held constant at a flat pitch.[15] Second, the trajectories from the accentual pitch

[14]Some of these constraints will be relaxed in future experiments. See §2.4.7 and §2.4.8 for experiments that relax the second and third constraints, respectively.

[15]There is mixed evidence on the relevance of the prenuclear pitch accent that is licensed on *Molly* in MAE (Chodroff & Cole, 2018; Im et al., 2023), where some work has described prenuclear accents as being primarily "ornamental," (Büring, 2016 though c.f., Petrone & Niebuhr, 2014 for Northern Standard German). However, in

targets to the ending pitch targets follow a linear trajectory. Lastly, the alignment of the accentual pitch target is held the same while the F0 value varies. The process for resynthesizing the F0 continuum[16] (using *Molly's from Branning* as an example) is as follows:

1. Resynthesize the prenuclear region (*Molly's from*) to be flat at a value of 90Hz.
2. Define a 5-step accentual pitch continuum centered on 90Hz. Each step of the continuum is spaced 10Hz apart, so the continuum spans from 70Hz (=L*) to 110Hz (=H*). Align these points to the end of the stressed syllable.
3. Define a 5-step ending pitch continuum within the speaker's range. This is operationalized in terms of F0 excursions from the low endpoint of the accentual pitch continuum, allowing for a slight fall from an L* (ERB scale differential: $-.25$ ERBs) and a large rise from an L* ($+2$ ERBs). Five equally spaced differentials between -.25 and +2 are created, then added to the value for L* (70Hz) to obtain five ending F0 targets. These points are aligned to the end of the phrase-final word (*Branning*).
4. Fully cross the two continua. Each contour is defined by an accentual pitch F0 target and an ending pitch F0 target. The trajectories between targets are linearly interpolated.

Figure 2.3 shows the time-normalized pitch contours of the resynthesized materials averaged across the five sentences.

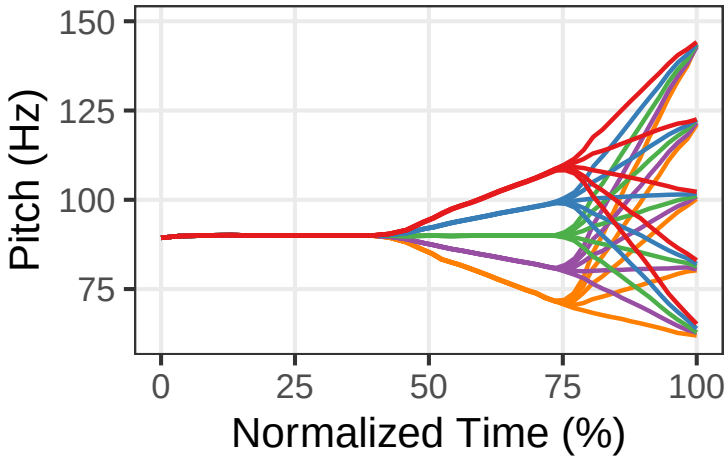


Figure 2.3: Exp. 1 (monotonal accents) materials. Time-normalized 25-step continuum across **the entire utterance**, including the prenuclear region, colored by accentual pitch target.

Castilian Spanish (a non-Germanic language), participants have been shown to be sensitive to prenuclear intonational patterns when judging interrogativity (Face, 2007).

[16]The values for the excursions are loosely based on the speaker's range from the original recordings as shown in Figure 2.1 and production results from Cole et al., 2023, who report speaker-normalized ERB-scale aggregate pitch contours for a variety target intonational tunes from an imitation task.

2.4.2 Procedure

Similar to Jeong (2018), this work uses a two-alternative forced choice (2AFC) task to probe the probabilistic association between accentual/ending pitch and ASSERTIVE or INQUISITIVE interpretations. A single trial of the task proceeds as follows. First, participants listen to one of the resynthesized audio files over headphones and are then asked to judge whether the speaker is **telling them something** or **asking them something**.[17] Participants make their selection using the F and J keys on their keyboard, where the mapping between the TELLING and ASKING response options is randomized by participant. The participant's selection is then recorded. To avoid order effects (Schiefer & Batliner, 1991) and the possibility that participants compare contours across trials via engagement in subvocal rehearsal (Baddeley, 2003; Jacquemot & Scott, 2006) of the stimuli, participants also count aloud by twos between each trial.

2.4.3 Predictions

In anticipation of the results, the predictions of the hypotheses will be described using a schematic version of the visualization that will be used when reporting our results. The results and predictions are depicted with a 5×5 heatmap, which shows the aggregate proportion of TELLING responses as accentual pitch and ending pitch are independently manipulated. Here, each stimulus pitch contour is a combination of one step from each continuum. Figure 2.4 shows the mapping between the 5×5 continuum and the 5×5 heatmap.

An effect of accentual pitch on the proportion of TELLING responses would be reflected by horizontal gradation along the heatmap while an effect of ending pitch would be reflected by vertical gradation along the heatmap. An interaction between the two would be reflected by additional

[17] This verbiage was selected for a few reasons. First, to maintain parallelism in the response option as opposed to "making a statement" (which has the additional sense of making a **political** statement) versus "asking a question." Second, to avoid "saying/asking something," as a speaker is always **saying** something when they speak. Lastly, to avoid linguistic jargon like "assertion" to avoid potential negative social connotations with "being assertive" (c.f. the formal sense of "proposing to add to the common ground" in pragmatics, Stalnaker, 1978). Note that Jeong (2018, p. 318) provided response options of *The speaker is giving out information/The speaker is seeking information* (in response to *What is the most likely interpretation of the utterance you heard?*) and *Oh, I didn't know that./Yes, didn't you know* (in response to *What is the more likely follow-up response tot he utterance you heard?*). The results from her experiments using these two different verbiages were comparable.

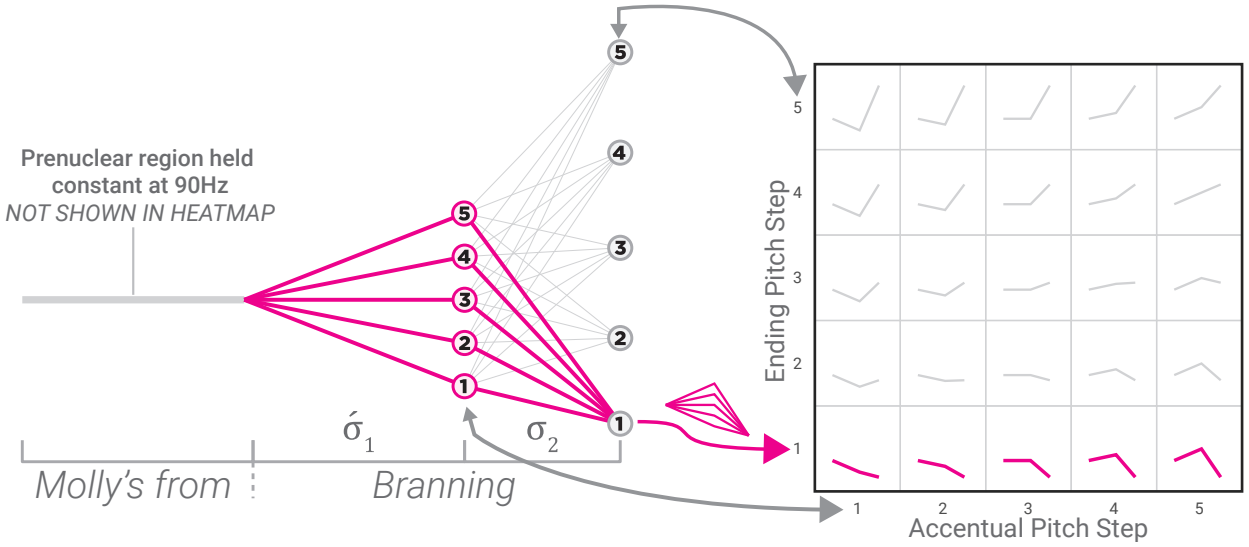


Figure 2.4: Mapping between the 5×5 continuum and the 5×5 heatmap used for the results, where pitch contours from the continuum are shown within each cell of the heatmap. Accentual pitch manipulations are shown on the X-axis. Ending pitch manipulations are shown on the Y-axis. **Only the pitch contour over the final word (*Branning*) is shown in the heatmap**—the constant prenuclear region is not shown. Note that because the accentual pitch target is aligned to the **end of the stressed syllable**, this manipulation appears roughly in the **middle** of the contour across *Branning*—**not** at the beginning of the contour.

modulation along the diagonal of the heatmap. Evidence for the Edge-only Hypothesis would be indicated by an effect of ending pitch only (vertical gradation), while evidence for the Integrative Hypothesis would be indicated by an effect of both ending pitch (vertical gradation) and accentual pitch (horizontal gradation). These patterns are tested using logistic regression to model the probability of a TELLING response, where a negative effect of ending pitch is predicted (a TELLING response is less likely with higher ending pitch) and potentially a positive effect of accentual pitch (a TELLING response is more likely when pitch accent is more like H*). When considering H*L-L% as the canonical intonation for assertions and L*H-H% as the canonical intonation for rising declaratives, we might find additional modulation in the quadrants that correspond to these tunes, which would be shown by an interaction between accentual and ending pitch. Figure 2.5 depicts what the predicted heatmaps would look like for each hypothesis.

All analyses are carried out using Bayesian mixed effect logistic regression models. Fixed effects of accentual pitch and ending pitch are included as continuous variables (as semitones

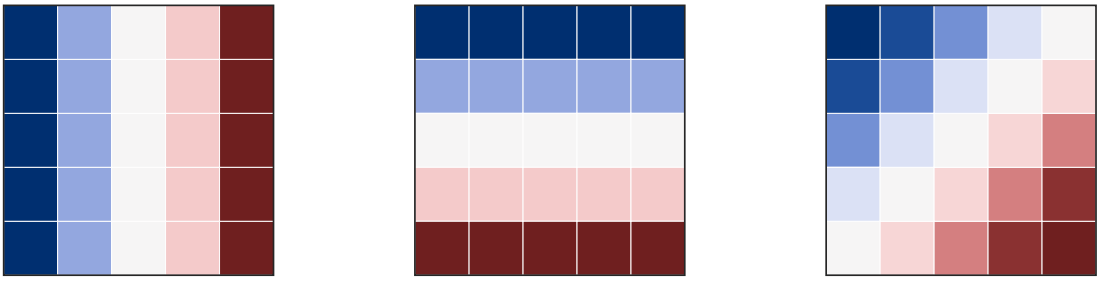


Figure 2.5: Schematic predictions for each hypothesis in terms of our heatmap, where red indicates higher proportions of ASSERTIVE interpretations and blue indicates lower proportions (white=50% either way). The left panel shows an effect of accentual pitch (=horizontal gradation). The middle panel shows an effect of ending pitch (=vertical gradation; this is the prediction for the edge-only hypothesis). The right panel shows an additive effect of both accentual and ending pitch (=diagonal gradation, this is the prediction for the integrative hypothesis).

from 90Hz, the accentual pitch continuum midpoint) along with an interaction between the two. Additionally, the model uses a random effects structure of random intercepts by participant and sentence (e.g., *Molly's from Branning*) and random slopes of accentual pitch, ending pitch, and their interaction by participant with the intuition that participants may be more or less sensitive to variation in one or more acoustic cues.

2.4.4 Participants

Participants for this experiment were recruited from Prolific (n=63), an online participant recruitment platform. Potential participants were initially screened using the platform-provided filters to recruit self-identified native English speakers that grew up in the United States. Participants were paid at a rate of $13.50/hour, commensurate with Evanston's minimum wage at the time data was collected. Seven participants were excluded,[18] leaving a total of 56 participants (31F, 24M, 1 Other, average age 37.3) for the analysis.

[18]The criteria for participant exclusion are the same across all experiments reported in this work. Specifically, participants were excluded if they self-reported any of the following: an uncorrected vision, reading, or hearing issue; if English is not their first language; or if they did not grow up in the United States.

2.4.5 Results

The aggregate proportion of TELLING responses from the empirical data is shown in Figure 2.6. From this figure, we can observe that there is strong vertical gradation and seemingly weaker horizontal gradation, where the proportion of TELLING responses is highest in the bottom-left (L*L-L%) quadrant. Curiously, the bottom-right (H*L-L%) quadrant is not at ceiling in the proportion of TELLING responses despite H*L-L% being the canonical intonational tune for assertions (Farkas & Bruce, 2010; Pierrehumbert & Hirschberg, 1990)— this point will be revisited when motivating the next experiment.

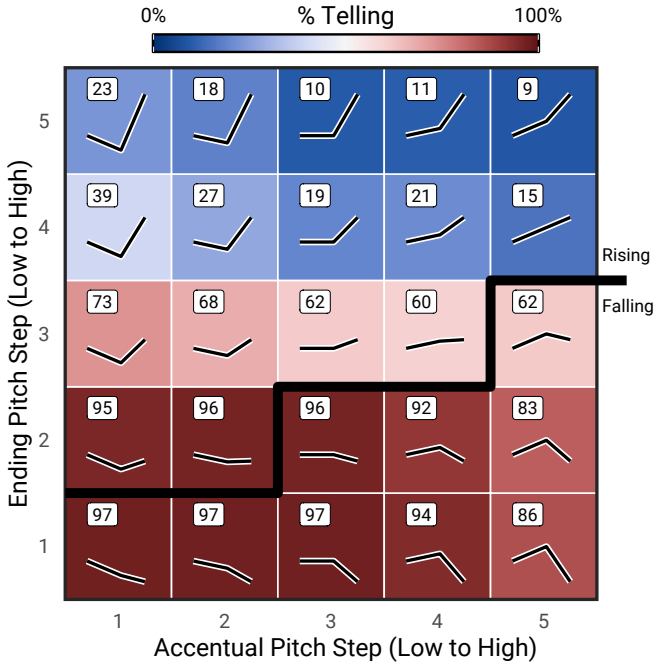


Figure 2.6: Exp. 1 (monotonal accents) empirical results. Aggregate proportion of TELLING responses for the pitch contours varying by accentual pitch on the X-axis (between/within monotonal L* and H*) and ending pitch on the Y-axis. The four corners (from left to right, top to bottom) correspond to nuclear tunes of L*H-H%, H*H-H%, L*L-L%, and H*L-L%.

The results of the Bayesian mixed effects logistic regression model is shown in Table 2.1, which reports the model estimates for each coefficient, their estimated standard errors, and the 95% credible interval (CrI). Evidence for the existence of an effect is taken to be reflected by the 95% CrI not containing 0.

Term	Estimate	Std.Error	95% CrI
Intercept	1.97	0.31	[1.35, 2.60]
AccentualPitch	−0.15	0.04	[−0.23, −0.06]
EndingPitch	−0.60	0.04	[−0.68, −0.52]
:AccentualPitch	0.00	0.01	[−0.01, 0.02]

Table 2.1: Logistic regression model results. Estimates are shown on the log-odds scale, where a higher likelihood of TELLING responses is reflected by positive values and a lower likelihood of TELLING responses (=higher likelihood of ASKING responses) is reflected by negative values.

The statistical model shows a positive intercept, reflecting an overall bias towards TELLING responses ($\hat{\beta} = 1.97, CrI = [1.35, 2.6]$). There is also strong evidence for an effect of ending pitch in the predicted direction ($\hat{\beta} = -0.6, CrI = [-0.68, -0.52]$) such that higher ending pitch (i.e., rising to higher F0 targets) is associated with a lower probability of a TELLING response. There is also evidence for an effect of accentual pitch, but not in the predicted direction ($\hat{\beta} = -0.15, CrI = [-0.23, -0.06]$) such that higher accentual pitch is also associated with a lower probability of a TELLING response. There is no evidence for an interaction between accentual pitch and ending pitch ($\hat{\beta} = 0, CrI = [-0.01, 0.02]$).

2.4.6 Discussion of Experiment 1

Experiment 1 manipulated accentual pitch between endpoints corresponding to the monotonal H* and L* pitch accents while separately manipulating ending pitch between high and low endpoints, yielding a variety of rising and falling contour shapes. There was a strong effect of ending pitch such that higher ending F0 targets were associated with a higher probability of ASKING responses. This finding differs from Jeong, 2018 in that a high ending F0 target yields a high probability of INQUISITIVE interpretations even when the pitch accent is H*, despite the fact that this pitch contour has a shallower slope compared to the pitch contour for canonical L*H-H%. In particular, the finding that participants were at-chance when judging Jeong's shallowest rise as asking or telling does not hold when looking at contours that more closely match H*H-H% (i.e., the top right quadrant of Fig. 2.6). Yet, the **negative** effect of accentual pitch is counter to the initial

prediction for this task. If H*L-L% is the canonical tune for assertions, why is it not at ceiling in terms of the proportion of TELLING responses? Relatedly, why do tunes more similar to L*H-H% show slightly higher proportions compared to H*H-H%?

With regard to the falls, one possible explanation for the unexpected effect of accentual pitch on TELLING responses may relate to the way continuous pitch contours were generated from the accentual pitch and ending pitch target F0 values. Following Jeong, the falling trajectories between the accentual and ending F0 targets were implemented via straight-line interpolation which yielded a gradual fall that sounded, admittedly, somewhat unnatural. In natural productions, pitch tends to falls more rapidly to a low F0 immediately following the accentual peak. As a result, it is possible that participants analyzed the gradually-falling contour differently than they would with a more natural "early fall," where F0 falls abruptly after reaching the accentual peak. So, the apparent effect of accentual pitch found in this experiment may be an artifact of the straight-line interpolation constraint. The next experiment relaxes this constraint.

2.4.7 Monotonal Pitch Accents with Early Falls (Exp. 2)

Experiment 2 builds on Experiment 1 to determine what the role of contour **shape** is for falling trajectories. The accentual pitch and ending pitch targets of the resynthesized materials are exactly the same as in Experiment 1, but now an additional low F0 target is added. This target is aligned at 30% of the second syllable duration and is equal in value to the ending pitch target. The averaged time-normalized pitch contours of the resynthesized materials are shown in Figure 2.7. Note here that the rises are exactly the same as in Experiment 1 and that the only difference is in the falls.

The prediction for this experiment is that the early fall trajectories should be closer to canonical H*L-L% trajectories, hence should be closer to ceiling compared to Exp. 1. As a result, the magnitude of the counterintuitive effect of accentual pitch should be reduced or eliminated altogether.

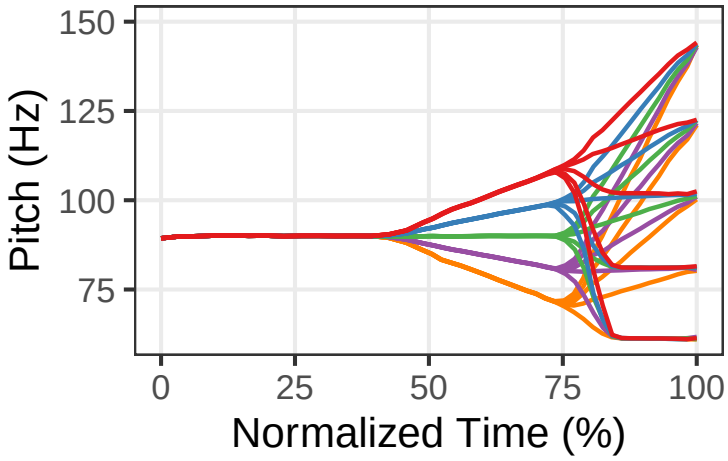


Figure 2.7: Exp. 2 (monotonal accents with early falls) materials. Time-normalized 25-step continuum across the entire utterance, colored by accentual pitch target. Compared to the contours in Fig. 2.3, the falling steps in this contour fall quickly, rather than gradually, to the low F0 target.

2.4.7.1 Results

A new group of participants who did not participate in Exp. 1 were recruited from Prolific (n=59). Again, seven participants were excluded using the same exclusion criteria as before, leaving a total of 52 participants (27F, 22M, 3 Other) for the analysis. The aggregate proportion of TELLING responses for the early fall continuum is shown in Figure 2.8.

From Figure 2.8, we can observe that the H*L-L% (bottom-right) quadrant has increased in the proportion of TELLING responses (relative to the results of Exp. 1) while the L*L-L% (bottom-left) quadrant remains at ceiling. This pattern is in line with the prediction for this experiment, but notably the pitch contours closest to H*L-L% are **still** not at ceiling like those closest to L*L-L%—this observation will be returned to in the discussion. The statistical model is exactly the same as the one used in the previous experiment, and the results are shown in Table 2.2.

The statistical model again finds an overall bias towards TELLING responses, as shown by the positive intercept ($\hat{\beta} = 2.36, CrI = [1.76, 2.95]$). The magnitude is greater than that of Experiment 1, which reflects the higher proportion of TELLING responses for the H*L-L% quadrant compared to Experiment 1. There is again a strong negative effect of ending pitch ($\hat{\beta} = -0.76, CrI = [-0.87, -0.66]$) and no evidence of an interaction between accentual pitch and

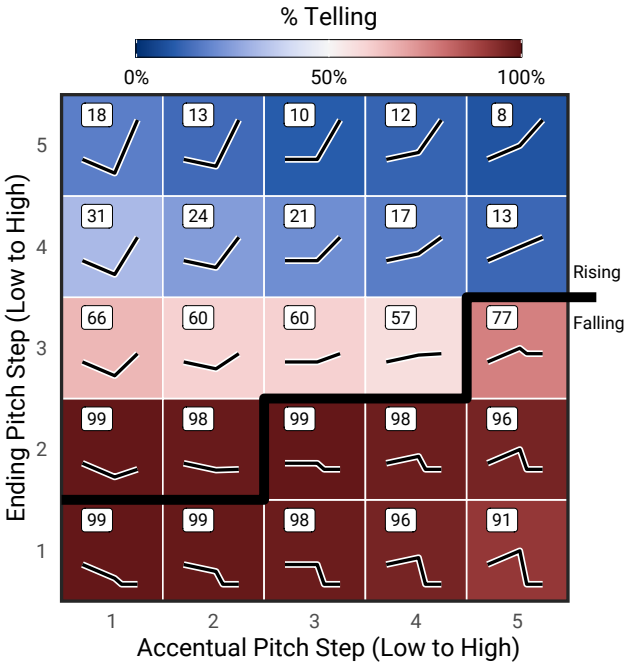


Figure 2.8: Exp. 2 (monotonal accents, early falls) aggregate proportion of TELLING responses. Pitch contours vary by accentual pitch on the X-axis (between/within monotonal L* and H*) and by ending pitch on the Y-axis. The four corners (from left to right, top to bottom) correspond to nuclear tunes of L*H-H%, H*H-H%, L*L-L%, and H*L-L%.

ending pitch ($\hat{\beta} = 0, CrI = [-0.02, 0.03]$). Importantly, there is now much weaker evidence for an effect of accentual pitch ($\hat{\beta} = -0.05, CrI = [-0.13, 0.03]$), though the bulk of the posterior distribution for the estimate is still on the negative side.[19]

2.4.7.2 Comparison with Longer Materials (Exp. 2b)

Recall that the straight-line interpolation constraint for the experimental materials was relaxed in Exp. 2, which nonetheless kept the shortening of the second syllable that was intended to ameliorate the oddity that came from the linear trajectories across the second syllable in the first place. Briefly, to verify that this shortening is not playing an adverse role in participants' interpretation of H*L-L% specifically, Exp. 2 was conducted once more using the full-duration resynthesized materials. All other aspects of the resynthesis procedure were kept exactly the same. So, the only difference here compared to Exp. 2 is that the second syllable length is now 364ms instead of

[19]The probability of direction (PD) of an effect, which ranges from 50% to 100%, is 90.76% for the effect of accentual pitch in Exp. 2. The PD for the effect of accentual pitch in Experiment 1 was 99.95%.

Term	Estimate	Std.Error	95% CrI
Intercept	2.36	0.30	[1.76, 2.95]
AccentualPitch	−0.05	0.04	[−0.13, 0.03]
EndingPitch	−0.76	0.06	[−0.87, −0.66]
:AccentualPitch	0.00	0.01	[−0.02, 0.03]

Table 2.2: Logistic regression model results for Experiment 2. Estimates are shown on the log-odds scale, where higher likelihood of TELLING responses is reflected by positive values and lower likelihood of TELLING responses (=higher likelihood of ASKING responses) is reflected by negative values.

255ms.

A new group of participants who did not participate in any of the previous experiments was recruited from Prolific (n=60). Four participants were excluded, leaving 56 participants (32F, 23M, 1 Other, average age 36.1) for the analysis. The results generally replicate the results of the previous experiments. Specifically, the statistical model finds a small negative effect of accentual pitch ($\hat{\beta} = -0.08, CrI = [-0.14, -0.01]$); this is again lower in magnitude than what was obtained in Experiment 1 ($\hat{\beta} = -0.15, CrI = [-0.23, -0.06]$) though not as low as what was found for Experiment 2 ($\hat{\beta} = -0.05, CrI = [-0.13, 0.03]$). There remains a strong negative effect of ending pitch ($\hat{\beta} = -0.6, CrI = [-0.69, -0.51]$) and small evidence of a positive interaction ($\hat{\beta} = 0.01, CrI = [0, 0.03]$), though this seems to be primarily driven by shallow rises becoming more like plateau trajectories (i.e., the middle row of the heatmap) as a result of the increased duration, thus increasing their proportion of TELLING responses.[20] The main finding from this experiment is that the duration manipulation does not substantially change the results; moving forward, the shorter syllable duration (as used in Exp. 1 and 2) will continue to be used.

2.4.8 Monotonal Pitch Accents with Varied Accent Alignment (Exp. 2c)

Experiment 2b focused on the observation that higher accentual pitch for falls appeared to decrease the proportion of TELLING responses, which was ameliorated through the early fall manipulation. This experiment follows up on the similar pattern in the rises, where L*H-H% appeared to be

[20] The heatmap and table of results are available in Appendix A.2.

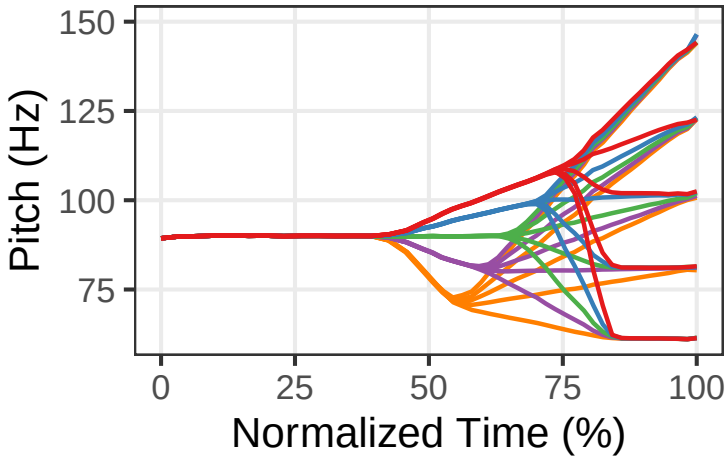


Figure 2.9: Exp. 2c (monotonal accents with varying alignment) materials. Time-normalized 25-step continuum across the entire utterance, colored by accentual pitch target. Compared to the contours in Fig. 2.7, the accentual pitch targets vary in both alignment and value.

less INQUISITIVE than H*H-H%. One reason this might be is because the L* target is normally expected to be aligned earlier to the stressed syllable compared to H*, but in the materials used so far the L* target is aligned to the very end of the stressed syllable. With reference to the fundamental parameters for rises and falls described in the introduction, this design was selected in order to vary only a single parameter (the "starting value") rather than two at once (both value and alignment). Here, this constraint is relaxed to vary the alignment of the accentual pitch targets such that the low end of the accentual pitch continuum (i.e., L*) occurs earlier in the stressed syllable.

The continuum of accentual pitch values, in terms of F0, are exactly the same as in the previous experiments. The alignment varies in equidistant steps from 30% of the stressed syllable (for L*) to 100% of the stressed syllable (for H*, matching the previous experiments). The resulting continuum is shown in Figure 2.9.

2.4.8.1 Results

A new group of participants who did not participant in any of the previous experiments was recruited from Prolific (n=60). Four participants were excluded, leaving a total of 56 participants (40F, 15M, 1 Other) for the analysis. The aggregate proportion of TELLING responses are shown

in Figure 2.10.

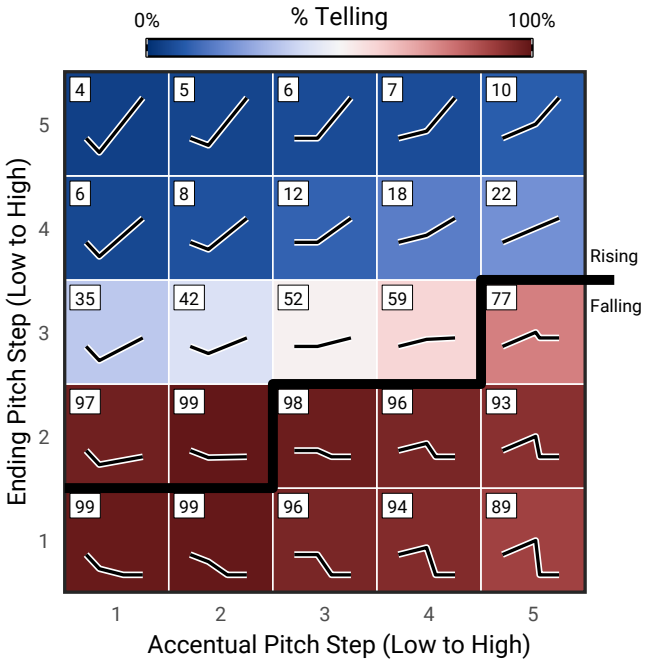


Figure 2.10: Exp. 2c (monotonal accents with varied alignment) aggregate proportion of TELLING responses. Pitch contours vary in the scaling **and alignment** of accentual pitch along the X-axis (between/within monotonal L* and H*) and in the scaling of ending pitch on the Y-axis. The four corners (from left to right, top to bottom) correspond to nuclear tunes of L*H-H%, H*H-H%, L*L-L%, and H*L-L%.

From Figure 2.10 we can observe that the alignment manipulation was successful in bringing down the counterintuitively high proportions of TELLING responses in the upper left quadrant. The results of the statistical model are shown in Table 2.3.

The statistical model again finds a strong credible effect of ending pitch ($\hat{\beta} = -0.86, CrI = [-0.99, -0.74]$), no credible effect of accentual pitch ($\hat{\beta} = 0.02, CrI = [-0.05, 0.09]$), and a small interaction between accentual pitch and ending pitch ($\hat{\beta} = 0.08, CrI = [0.05, 0.1]$). The interaction largely serves to account for the slightly higher proportions in the H*H-H% quadrant compared to the L*H-H% quadrant.

Term	Estimate	Std.Error	95% CrI
Intercept	1.90	0.24	[1.42, 2.39]
AccentualPitch	0.02	0.04	[−0.05, 0.09]
EndingPitch	−0.86	0.06	[−0.99, −0.74]
:AccentualPitch	0.08	0.01	[0.05, 0.10]

Table 2.3: Logistic regression model results for Experiment 2c. Estimates are shown on the log-odds scale, where higher likelihood of TELLING responses is reflected by positive values and lower likelihood of TELLING responses (=higher likelihood of ASKING responses) is reflected by negative values.

2.5 Interim Discussion of Monotonal Pitch Accent Experiments

Experiments 1 through 2c used rising and falling contours with monotonal pitch accents that progressively relaxed constraints imposed on the materials. Experiment 2 found that using the early fall trajectories increased the proportion of TELLING responses for all falling steps of the continuum, leading to a reduction in the magnitude of the effect of accentual pitch. Moreover, this reduction in magnitude holds (albeit to a smaller degree) even when revisiting the durational manipulation made in Experiment 2b to ameliorate the materials in Experiment 1. When considering the rising steps of the continuum, Experiment 2c reversed the counterintuitive pattern found in previous results (where the % TELLING was higher for L*H-H% than H*H-H%) by shifting the alignment of the L* target earlier. Across the four experiments, a robust main effect of ending pitch was found such that the proportion of TELLING responses decreases drastically as ending pitch increases.

Thus far, the results point towards ending pitch—that is, the choice and scaling of the boundary tone—as the primary cue for Q/A interpretation for rises and falls. When considering manipulation of other parameters, such as accentual pitch value and alignment, the results are more mixed. The results suggest that the more natural early alignment of L* is more characterstic of INQUISITIVE rising declaratives, while later alignment is less so. Three observations from the results motivate this. First, Experiment 2c shows a small positive effect of accentual pitch which is in line with Jeong's proposal linking H*H-H% to ASSERTIVE rising declaratives. However, the contours clos-

est to H*H-H% (in the top right quadrant of each heatmap) are still solidly on the INQUISITIVE side, with proportions lower than what was observed in Jeong's results (there, a range of ≈34%-47% and here, a range of 8%-22%). Second, The contours closest to Jeong's pattern of results are in the middle row of each heatmap (range for the first three columns in the middle row: 60%-73% in Exp. 1 and 2 and 35%-52% in 2c), with mid-level ending pitch targets. Third, When considering the contours closest to L*H-H% (in the top left quadrant of each heatmap), the earlier-aligned L* targets showed a range of 4%-8% while the later-aligned contours showed a range of 13%-39%. Taking these results together, we see that variation in the height of the accentual pitch target appears to play only a limited role in ASSERTIVE interpretations whereas the height of the ending F0 target plays a robust role in making INQUISITIVE interpretations more likely. Based on the alignment results and how lower ending F0 targets increased the proportion of TELLING responses, it appears that deviating from the canonical form of L*H-H% makes INQUISITIVE interpretations less likely—but H*H-H% is not the only way in which this deviation can be made.[21]

2.5.1 Pattern in the Falls

While the present work has been largely motivated by work on rising intonation and expanding the range of materials through manipulating different fundamental parameters, it is worthwhile to see how variation in these parameters extends to falls. Notably, across the heatmaps, falls show much less variation and overwhelmingly receive TELLING interpretations. There is a notable pattern that is persistent across the experiments presented thus far: Higher accentual pitch appears to be associated with lower, not higher, proportions of TELLING responses for falling contours. At first glance, one may be tempted to attribute this to a lowered ceiling in the experiment where the

[21] One might object that it is not typical for L* to be aligned so late, so the deviation from L*H-H% may be outside the bounds of what would be naturally expected. One way to reconcile this may be to recast the "late-aligned" L*H-H% steps as something like H+!H*H-H%, with a falling onglide towards an accentual F0 target lower than the prenuclear region. However, the point is that deviating from a canonical L*H-H%, characterized by a low early-aligned accentual target and a high ending F0 target, can be made even without these steps, as the H*H-H% quadrants show small increases in proportions of TELLING interpretations while the middle rows of the heatmaps show more substantial increases. If the results were solely driven by H*H-H% being linked to ASSERTIVE rising declaratives, then the H*H-H% quadrant should show proportions close to chance to match the results of Jeong (2018, p. 320), but this pattern was not obtained.

highest proportions we receive are not 100% but rather are closer to 90% (e.g., 94% may be practically equivalent to ceiling if there is noise in the data arising from mistaken button presses). Such an account does not follow, however, as other falling contours with lower accentual pitch receive proportions as high as 99%. What is it about raised accentual pitch that makes falls systematically receive around 10% fewer TELLING responses?

Recall from the discussion of bitonal pitch accents that increased prominence can serve to prosodically mark focus (Rooth, 1992), where prominence displays continuous variation with graded but nonetheless overlapping distributions across accent types (Im et al., 2023; Watson, 2010). Accordingly, for falls, raising the accentual pitch as done in Experiments 1-2c will raise the prominence of the accented item and may, in turn, increase the likelihood that alternatives are relevant to the interpretation of the speaker's utterance. In the context of the task used here, which restricts the response options to TELLING or ASKING, it is possible that distinct interpretations, plausibly related to focus, are salient to the participant and detract from the Q/A contrast probed by the task.

While contrastive focus is one example of focus, focus can can serve many different communicative purposes such as CORRECTING or CONFIRMING information (Krifka, 2008, pp. 250–253) depending on the discourse context and the speaker's intended meaning, as in (1a and 1b). The higher prosodic prominence used to mark focus may also convey MIRATIVE FOCUS (1c), defined by Cruschina (2021, p. 2) as expressing "unexpectedness with respect to more likely alternatives," which can also be paraphrased as expressing SURPRISE or INCREDULITY (see also Rett & Sturman, 2020). Moving forward, the collection of such interpretations—CONTRASTING, CORRECTING, CONFIRMING, MIRATIVITY, etc.— will be referred to as FOCUS-ENRICHED INTERPRETATIONS.

(1) a. A: Is Molly from Skokie?

B: Molly's from [BRANNING]$_F$. (...**not** Skokie) CORRECTING

b. A: Molly's from Branning?

B: (Yes,) Molly's from [BRANNING]$_F$. CONFIRMING

c. A: Molly's from Branning.

B: (Wow) Molly's from [BRANNING]$_F$. (Of all places!) MIRATIVE

The experimental design used in this work is limited in the sense that, unlike the examples in (1), which can differ in the inferred context, no discourse context was provided to the participant. So, if such focus-enrichment occurs, the specific communicative function that the participant infers is relatively unconstrained, as in (1). That is, participants might imagine any number of contexts (as in 1) to accommodate such focus-enriched interpretations (see also discussion in Cole, 2015). It is important to note that such focus-enriched interpretations are, theoretically, not mutually exclusive to ASKING and TELLING. For instance, if you correct someone, you're still telling them some information. However, in terms of the social communicative goal that the speaker is achieving, the participant may nonetheless view things like CORRECTING as distinct from merely ASKING or TELLING. Put in other terms, what may be most salient to the participant is not that the speaker is merely telling them something, but that the speaker is making use of focus to correct some mistaken belief (or to confirm information, express their surprise, etc.). For example, if the participant interprets a contour as TELLING+CONTRAST, they could either respond with TELLING **or** seek out a response option that distinguishes CONTRAST from merely TELLING— but here the only non-TELLING option is ASKING.[22] Under this view, the participant's choice of ASKING in the experiment is the result of a task-specific response behavior which (for falls) treats ASKING as an ad-hoc "not-TELLING" option. In other words, the conjectured focus-enriched interpretation may be "interfering" with participants' response along the Q/A dimension, resulting in a reduction of TELLING responses for falls.

The proposal presented in this section is a post-hoc behavioral hypothesis seeking to account for the observed pattern that higher accentual pitch for the falls is associated with lower, not higher,

[22]It is assumed in this work that the expected interpretation for falls here is TELLING following work from Gunlogson (2001, p. 1) (among many others) describing that declarative sentences with falling intonation is "the canonical way to make a statement" vis-à-vis rising declaratives where "the rise seems to impart the force of a question to what would otherwise be naturally interpreted as a statement." Additionally, based on the statistical results presented thus far, where positive intercepts denoted a bias toward TELLING responses, it appears uncontroversial to expect that the "default" response for falls should be TELLING—hence why the focus here is on the association between raised accentual pitch and lower proportions of TELLING responses.

proportions of TELLING responses in these experiments. Because accentual prominence is known to be associated with prosodic focus for falling intonation in MAE, this naturally leads one to seek out an explanation for the pattern in terms of focus. Yet, focus itself is orthogonal to assertions and questions: One can invoke alternatives to a constituent when making an assertion or asking a question.[23] So, the proposal here should not be taken to suggest that the marking of focus is inherently inquisitive; rather, the point is to account for a counterintuitive pattern resulting from limitations of the paradigm. If it is the case that prosodic prominence is associated with increased salience of other interpretations which, in turn, interfere with participants' response behavior in the task, then we should expect that further increasing prominence would make this pattern more apparent. This prediction motivates the next experiment, which uses a more prominent pitch accent that is commonly associated with (if not equated with) prosodic focus.

2.6 Bitonal Pitch Accent Scaling (Exp. 3)

Experiments 1-2c investigated various acoustic manipulations with monotonal pitch accents and so they are not able to comment upon the broader pitch accent inventory in MAE. The next experiment continues to probe whether the pitch accent of the contour plays a role in interpretation of the Q/A contrast by investigating whether a bitonal L+H* makes additional interfering interpretations more likely, thus reducing the proportion of TELLING responses. Prior work shows that L+H* is generally more prominent than H* (i.a. Im et al., 2023), and so if competing focus-enriched interpretations are more likely for falls when accentual prominence is increased, then the use of a more prominent L+H* accent should further increase the likelihood of such interpretations. Given the results from the previous experiments, the prediction for the falls is straightforward: Increased prominence should increase the likelihood of focus-enriched interpretations, which interfere with judgments along the Q/A contrast, lowering the proportion of TELLING responses for falls. The L+H* accent is not frequently discussed in conjunction with the H-H% edge-tone configuration,

[23] Although focus is related to questions in the sense of question-answer congruence, e.g., an answer of "Molly's from BRANNING." is felicitous in response to "Where is Molly from?" but not to "Who's from Branning?", this appears unrelated to the present set of results. One might wonder how the results may differ if such a question manipulation was added, but this is beyond the scope of the current work.

so the prediction for Q/A interpretation of rises with L+H* is not obvious and will be somewhat exploratory in this experiment.

2.6.1 Materials

At a high level, this experiment investigates within-category scaling of the L+H* pitch accent. This rising pitch accent has two primary acoustic correlates. First, it contains a low tonal target prior to the high accentual peak, which leads to a more extreme pitch excursion in the rising onglide across the accented syllable. Second, while standard AM theory would describe this rising onglide in terms of linear interpolation between the two tonal targets (L+ and H*), it has been observed that L+H* has a somewhat domed trajectory (Barnes et al., 2021; Iskarous et al., 2024; Steffman & Cole, 2024; Steffman et al., 2024). Accordingly, the materials include a low target early in the accented syllable and implement a curved, dome-like, trajectory using Bézier curves (details are described in Appendix A.3). To ensure that **all** steps along the accentual pitch continuum qualify as **rising** pitch accents from a 70Hz L+ target, the continuum is shifted up by 10Hz (relative to the continua used in Experiments 1 and 2) from a range of [70,110] to [80,120]. The ending pitch values are unchanged from the previous experiments and use the early-fall trajectories previously introduced in Exp. 2b. The averaged resynthesized contours are shown in Figure 2.11.

2.6.2 Results

The procedure is the same as in the previous experiments. A new group of participants who did not participate in any of the previous experiments were recruited from Prolific (n=59). Four participants were excluded, leaving a total of 55 participants (30F, 22M, 3 Other, average age 38.5) for the analysis. Recall that if perceived higher prominence licenses a focus-enriched interpretation, which in turn detracts from the total proportion of TELLING responses, then we should expect lower proportions of TELLING responses for falling contours closest to L+H*L-L% (bottom right). The aggregate proportion of TELLING responses is shown in Figure 2.12.

From Figure 2.12, we can observe strong gradation for the falling steps of the continuum as ac-

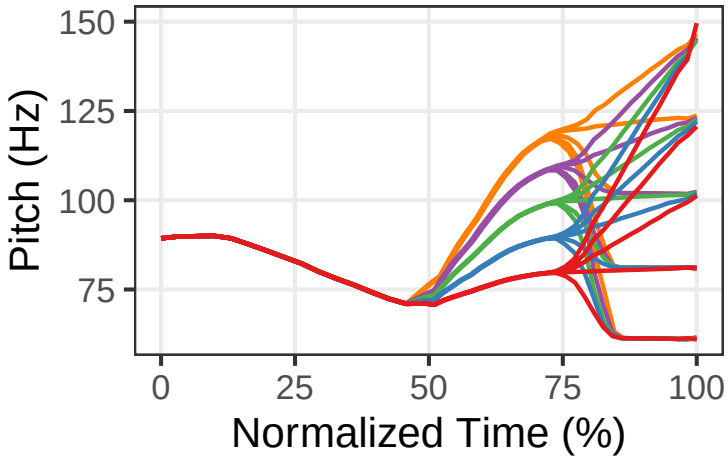


Figure 2.11: Exp. 3 (L+H* scaling) materials. Time-normalized 25-step L+H* continuum across the entire utterance, colored by accentual pitch target. Compared to the contours in Fig. 2.7, the accentual pitch targets are shifted up by 10Hz and have a domed onglide from a low F0 target (70Hz) aligned to the start of the nuclear word.

centual pitch is scaled higher. This observation is in line with the prediction that the proportion of TELLING responses should decrease for falls with higher-scaled L+H* accents under the hypothesis that salient focus-enriched interpretations interfere with participants' judgments. With regard to the rising contours, we can observe that the proportions of TELLING responses have generally increased relative to Exp. 2 (the range for the top two rows in Exp. 3 is [15,37] c.f. [8,31] in Exp. 2 and [4,22] in Exp. 2c) but that there is no visually apparent gradation akin to what is seen with the falls. In other words, the scaling of the L+H* accent does not strongly influence Q/A interpretation for rising contours.

Table 2.4 shows the results of the statistical model following the same structure as in previous experiments. From the model we find, once again, a strong effect of ending pitch ($\hat{\beta} = -0.52, CrI = [-0.6, -0.45]$) and no evidence for a main effect of accentual pitch ($\hat{\beta} = -0.03, CrI = [-0.13, 0.07]$). However, there is evidence of a small interaction between the two ($\hat{\beta} = 0.03, CrI = [0.02, 0.05]$). Notably, there is a positive intercept ($\hat{\beta} = 1.73, CrI = [1.2, 2.25]$) that is lower in magnitude than that of Experiment 2 ($\hat{\beta} = 2.36, CrI = [1.76, 2.95]$). The reduction in magnitude of the intercept is likely to be primarily driven be the reduced proportions of TELLING responses observed for the falling steps.

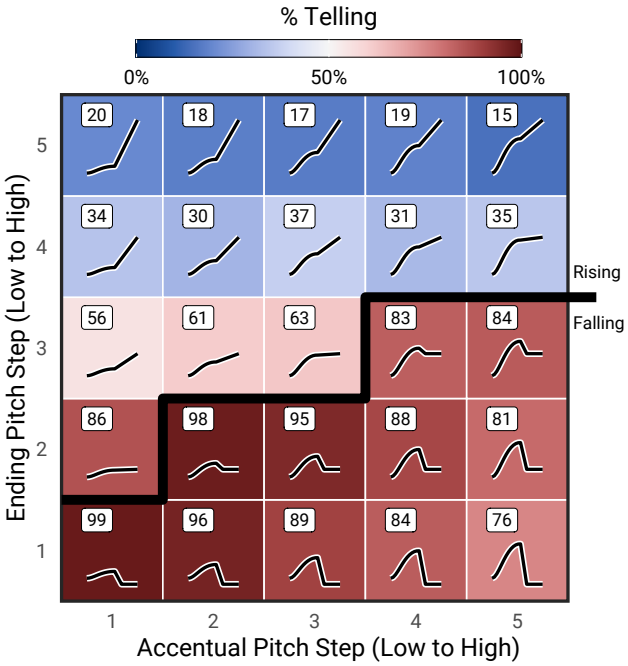


Figure 2.12: Exp. 3 (L+H* Scaling) aggregate proportion of TELLING responses. Pitch contours varying in the scaling of accentual pitch on the X-axis (within bitonal L+H*) and in the scaling of ending pitch on the Y-axis.

2.6.2.1 Post-hoc Subset Analysis

One limitation to the analysis for this set of results is that the fixed effects of the statistical model merely test the hypotheses that the pitch accent and edge-tone configuration play a "global" role in interpretation (i.e., the effect is the same across both rises and falls). In this set of results, scaling of the L+H* accent does not appear to play a role in the rising steps, in contrast to the rather obvious variation in the falls, hence precluding a global effect. Thus, despite our main observation being the

Term	**Estimate**	**Std.Error**	**95% CrI**
Intercept	1.73	0.27	[1.20, 2.25]
AccentualPitch	−0.03	0.05	[−0.13, 0.07]
EndingPitch	−0.52	0.04	[−0.60, −0.45]
:AccentualPitch	0.03	0.01	[0.02, 0.05]

Table 2.4: Logistic regression model results. Estimates are shown on the log-odds scale, where higher likelihood of TELLING responses is reflected by positive values and lower likelihood of TELLING responses (=higher likelihood of ASKING responses) is reflected by negative values.

gradation seen in the **falls**, this model is unable to narrowly identify such a role of the pitch accent that is conditioned on the global shape. To gain a better picture of the potential conditional effect of bitonal accent scaling, a second statistical model is fit to a subset of the data. With reference to the heatmap of empirical results, the model uses the bottom two rows as falling trajectories and the top two rows as rising trajectories. The middle row, as it is split between falls and rises with excursions that are shallow enough to be more similar to plateaus, is omitted.

The logistic regression model is parameterized using accentual pitch scaling as a continuous variable (in the same way as in previous models) as it interacts with a two-level categorical variable of global shape, with falling as the reference level (coded as 0) and rising as the comparison level (coded as 1). This contrast coding scheme gives us the at-issue conditional effect of accentual pitch for falls (i.e., testing directly for the horizontal gradation among falls). Based on the results shown in Figure 2.12, there is not a strong pattern of gradation among the top two rows, so we would expect the conditional effect of accentual pitch for rises to be close to zero. The results of this model[24] are shown in Table 2.5.

Term	**Estimate**	**Std.Error**	**95% CrI**
Intercept	2.85	0.27	[2.31, 3.38]
AccentualPitch	-0.22	0.05	$[-0.32, -0.12]$
RisingShape	-3.72	0.28	$[-4.24, -3.14]$
:AccentualPitch	0.18	0.06	[0.07, 0.29]

Table 2.5: Logistic regression model results. Estimates are shown on the log-odds scale, where higher likelihood of TELLING responses is reflected by positive values and lower likelihood of TELLING responses (=higher likelihood of ASKING responses) is reflected by negative values.

The new statistical model finds the expected negative conditional effect of accentual pitch scaling for falls ($\hat{\beta} = -0.22, CrI = [-0.32, -0.12]$). This result straightforwardly reflects the predicted effect of scaling for the falling contours: higher prominence reduces the proportion of TELLING responses. The model also finds a positive interaction of accentual pitch and global

[24]There are other ways to parameterize this model, such as by treating boundary step as a five-level factor and identifying the effect of accentual pitch scaling for each row, or by including the middle row, or by splitting the data based on the dividing line. Regardless of the parameterization, the results are largely the same. However, different approaches are more or less sensitive to the endpoints and plateau-like contours in the middle, and so come with the caveat that some observed effects are largely driven by a corner of the heatmap.

rise/fall pattern which reflects how different the conditional effect of accentual pitch for the rises is compared to the falls. Here, the interaction is positive and roughly of equal magnitude to the conditional effect for the falls ($\hat{\beta} = 0.18, CrI = [0.07, 0.29]$), so variation in accentual pitch scaling for the rises is overall negative but very small in magnitude.[25] These patterns are reflected in Figure 2.13, where again the main takeaway is that high prominence for the falls yields lower proportions of TELLING responses while the proportions for the rises change very little.

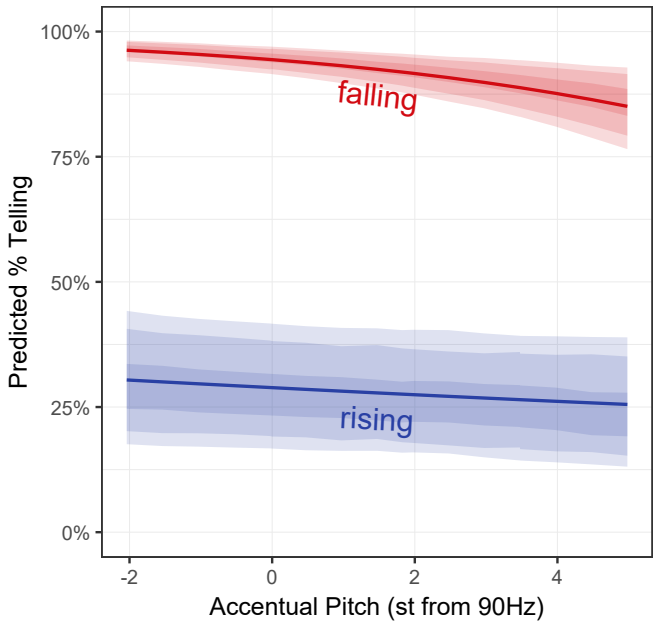


Figure 2.13: Exp. 3 (L+H* Scaling) posterior identification curves of the conditional effect of accentual pitch scaling for rising and falling contours with 50%, 89%, and 95% credible intervals.

2.6.3 Discussion of Scaling Results

Experiment 3 shifted from investigating variation between two monotonal pitch accents, L* and H*, to investigating within-category variation of a single bitonal pitch accent, L+H*. The results showed that that increasing the accentual peak of the L+H* pitch accent decreases the proportion of TELLING responses for the falling steps, in line with the prediction that increased prominence may make competing focus-enriched interpretations more likely, hence lowering the proportion of

[25] It is crucial to understand that this interaction estimate does **not** reflect a positive effect of accentual pitch for rises, as it needs to be added to the effect estimate for the falls (i.e., $-0.22 + 0.18 = -0.04$).

TELLING responses for the falls. Lastly, it was found that variation in the scaling of the accentual peak for L+H* did not play a large role in the interpretation of rises.

In some ways, the results of Experiment 3 are not particularly unexpected. L+H* is often referred to as the "focus-marking" or "contrastive" accent in the literature, so it is unsurprising to find that focus appears relevant to this set of results. What should be emphasized is that the results of this experiment lends credence to the hypothesis that increased prominence is associated with lower proportions of TELLING responses for the falls due to the increased availability of focus-enriched interpretations. However, whether these kinds of focus-enriched interpretations are actually salient to participants in this task is still yet to be directly addressed experimentally. Looking ahead a bit, Experiment 5 will more explicitly probe whether focus-enriched interpretations like those previously described are salient to participants. But, before then, there is one final set of manipulations to investigate.

When relating Experiments 1-3 back to the space of fundamental parameters relevant for rises and falls, we have so far investigated **scaling** in two regions (accentual and ending pitch), contour **shape** in two regions (the onglide to the accentual peak and the trajectory of falling contours), and the role of L* alignment (in Experiment 2c versus 1-2b). Still outstanding is the the role of accentual peak **alignment**, which is particularly important when describing the distinction between the L+H* and L*+H pitch accents and forms the basis of the next experiment.

2.7 Bitonal Pitch Accent Alignment (Exp. 4)

The L*+H accent is typically described as a rising bitonal pitch accent whose accentual peak is aligned relatively **late**, typically outside of the stressed syllable, in contrast to L+H* whose accentual peak is aligned **earlier** (Barnes et al., 2021). In other words, while the accentual peak of L+H* is (typically) reached at, or slightly before, the end of the stressed syllable, the accentual peak of L*+H is typically reached **after** the stressed syllable.[26] Alignment, however, is not the only difference between L+H* and L*+H. A frequent observation between the two pitch accents is

[26]This is of course assuming a second syllable exists after the stressed syllable, as in our materials. We set aside questions of tonal compression with monosyllabic words.

that L+H*, as previously mentioned, has a DOMED onglide while L*+H has a SCOOPED onglide.

The category-level distinction between L+H* and L*+H is somewhat contentious (and similarly so for H* and L+H*; see Ladd, 2022 for a recent discussion), with alternative proposals suggesting that the distinction relates to variants along a continuum of possible alignment values (Barnes et al., 2021; Gussenhoven, 1984). Work focusing on rising pitch accent alignment in the context of specific edge tone configurations has suggested evidence for categoricity, with Pierrehumbert & Steele (1989a) finding a bimodal distribution of accentual peak alignment for productions of L+H* and L*+H in the context of L-H% edge tones. Similarly, Dilley & Brown (2007) and Dilley & Heffner (2013) report a bimodal distribution of accentual peaks for bitonal accents in the context of following L-L% edge tones (with weaker evidence for bimodal valley alignment in the context of H-H%). However, recent work reported in Steffman et al. (2024) has taken a more exhaustive look at the putative contrast between rising accent categories (H*, L+H*, L*+H) using imitation and perceptual discrimination tasks, finding only small differences among the rising pitch accents with only some edge tones. They conclude that "the evidence for H* and L+H* as robustly distinct pitch accents is not strong, nor is the evidence for L+H* versus L*+H," [*ibid* p.25] particularly in comparison to the robust contrast between H* and L*. In other words, the distinction between L+H* and L*+H may reflect within-category variation rather than between-category variation.

The goal of the next experiment is to investigate the alignment contrast between the two rising bitonal pitch accents. The previous experiment showed that increasing the accentual peak scaling of L+H* reduced the proportion of TELLING responses for falling contours, but did not strongly affect rising contours. In this experiment, both bitonal accents are high on the prominence hierarchy described by Im et al. (2023), and thus both may license focus-enriched interpretations (at least, when followed by L-L% edge tones). Likewise, other descriptions of the meaning contributions for L+H* and L*+H have suggested that the two are similar, sharing the evocation of a scale or the marking of focus (Göbel, 2019; Pierrehumbert & Hirschberg, 1990). Based on these prior mixed results, the predictions for this experiment are weaker compared to those in Experiments

1-3, making Experiment 4 more exploratory than the preceding experiments. On the one hand, it may be that L*+H will be perceived as more prominent than L+H*, leading to further reduced proportions of TELLING responses for the falling steps. On the other hand, it could be that using either bitonal accent makes focus-enriched interpretations likely for falls, but that the alignment of the accent does not further modulate the results beyond what was seen in Exp. 3.

2.7.1 Materials

As with the materials for the previous experiment, Bézier curves (described in Appendix A.3) were used to create domed and scooped onglide trajectories. Unlike the previous experiments, however, the focus is on manipulating the alignment of the pitch accent targets (the initial low F0 target and the high accentual peak) rather than the scaling of the peak, which here is kept constant at a high target of 120Hz. The alignment continuum consists of equally-spaced points between endpoints of 80% and 115% of the stressed syllable duration.[27] The distance between the initial low target in the bitonal accent and the high accentual peak target is held constant for all steps of the continuum (i.e., both are shifted earlier/later, not just the peak target). Finally, the degree of curvature for the rising onglide is simultaneously manipulated with the alignment such that the later-aligned steps of the continuum are also more "scooped" and the early-aligned steps of the continuum are more "domed." The averaged resynthesized contours are shown in Figure 2.14.

2.7.2 Results

A new group of participants who did not participate in any of the previous experiments was recruited from Prolific (n=60). Two participants were excluded, leaving a total of 58 participants (28F, 28M, 2 Other, average age 39.1) available for analysis. The results are shown in Figure 2.15.

Based on Figure 2.15, we can observe that there is no apparent monotonic pattern of gradation for the falling trajectories. However, for the rising trajectories in row 4, we can observe that earlier alignment yields higher proportions of TELLING responses compared to later alignment.

[27] The endpoints were decided based on a gated procedure of decrementing/incrementing the alignment by 5% until either the result sounded too unnatural or the resynthesis quality was compromised.

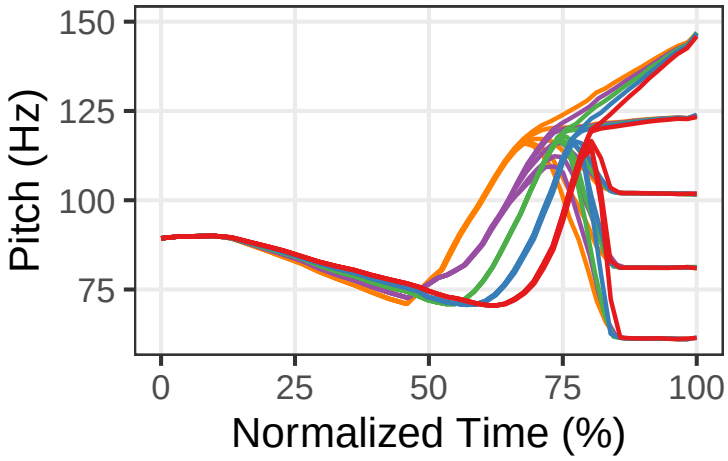


Figure 2.14: Exp. 4 (Bitonal Alignment) materials. Time-normalized 25-step continuum across the entire utterance, colored by accentual pitch target. Compared to the contours in Fig. 2.11, the accentual pitch targets are shifted in terms of alignment and shape but held at the same accentual F0 target (120Hz).

These results are again modeled using a Bayesian mixed-effects logistic regression model. Note that because the scaling of accentual pitch is no longer manipulated, the accentual pitch term in the model is replaced with a continuous variable corresponding to the alignment manipulation. This alignment predictor is centered on 100% (i.e., aligned exactly to the end of the stressed syllable) and is scaled such that a one-unit increase corresponds to a 10 percentage point increase in alignment.[28] The statistical model results are shown in Table 2.6.

Term	**Estimate**	**Std.Error**	**95% CrI**
Intercept	2.34	0.29	[1.79, 2.92]
Alignment	−0.03	0.07	[−0.16, 0.10]
EndingPitch	−0.55	0.05	[−0.66, −0.45]
:Alignment	−0.07	0.01	[−0.10, −0.04]

Table 2.6: Logistic regression model results for Exp. 4 (bitonal alignment). Estimates are shown on the log-odds scale, where higher likelihood of TELLING responses is reflected by positive values and lower likelihood of TELLING responses (=higher likelihood of ASKING responses) is reflected by negative values.

The statistical model once again finds a strong main effect of ending pitch ($\hat{\beta} = -0.55, CrI = [-0.66, -0.45]$). The model does not show evidence of a main effect of accent alignment ($\hat{\beta} =$

[28] For example, a positive effect would indicate that increasing alignment from 100% to 110% of the stressed syllable duration (i.e., later alignment) is associated with an increased likelihood of TELLING responses.

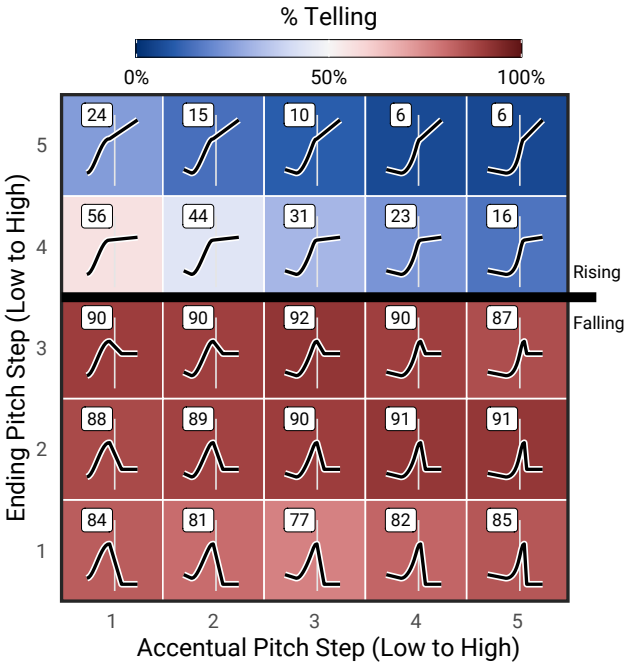


Figure 2.15: Exp. 4 (Bitonal Alignment) aggregate proportion of TELLING responses. Pitch contours vary in the **alignment** of accentual pitch on the X-axis (between early-aligned L+H* and late-aligned L*+H) and in the scaling of ending pitch on the Y-axis. Each cell shows a vertical line denoting the end of the accented syllable to better differentiate between (earlier/later) alignment steps.

$-0.03, CrI = [-0.16, 0.1]$). However, there is a negative interaction between alignment and ending pitch ($\hat{\beta} = -0.07, CrI = [-0.1, -0.04]$). This interaction primarily accounts for the gradation among the rising trajectories.[29]

2.7.3 Discussion

Experiment 4 investigated the effect of bitonal pitch accent alignment on Q/A interpretation. Generally, there were no systematic patterns in interpretation for the falling steps, suggesting that while the use of a bitonal accent may make alternative interpretations more salient, the choice of accent based on alignment does not further modulate this effect. For the rises, there appears to be more substantial variation.

[29]For parity with the post-hoc analysis in the previous experiment, a supplementary model modeling the conditional effect of alignment for rises and falls is fit. This supplementary model shows no conditional effect of alignment for the falls ($\hat{\beta} = -0.3, CrI = [-0.39, -0.21]$) but does show a strong negative effect for rises ($\hat{\beta} = -0.61, CrI = [-0.75, -0.47]$).

With regard to the rising contours in row 4, we can see that the pitch excursions from the accentual peak to the ending pitch are positive (rising) but nonetheless very small. Here, the gradation observed in the figure likely has less to do with the contribution of an L+H* or L*+H pitch accent and more to do with the duration over which the small positive final pitch excursion takes place. When the accentual peak is earlier, then the duration over which the (shallow) rise occurs is greater, greatly reducing the slope of the rise. As a result, these shallow rises are likely perceived more like a plateau (indicative of an H-L% edge-tone configuration). When the accentual peak is aligned progressively later the proportion of TELLING responses decreases. For these contours, there may be insufficient time[30] for the plateau trajectory to be implemented, rendering little evidence against an overall rising shape. When the ending F0 target is raised (moving from row 4 to 5), however, this still greatly increases the likelihood of an ASKING interpretation regardless of the alignment.

In terms of the original hypotheses, there is still see a stark divide between the falling and rising groups, indicating further support for the Edge-Only Hypothesis. The effect of alignment only makes an impact in the subset of materials where the rising pitch excursion is very small, becoming more plateau-like when the excursion takes place over a longer timespan. While plateaus are not associated with prosodic focus-marking, the interpretation of these results appears similar to the previous discussion of patterns in the falling contours: the plateaus may be more likely to be interpreted as if the speaker were LISTING multiple options. The speaker could be listing to question or assert multiple things, but the disambiguating final turn from the speaker never comes. While this LISTING function can be seen as distinct from ASKING or TELLING, it also is not straightforwardly an "enriched" meaning of either response option. As such, it is not surprising to see participants respond close to chance levels for plateaus.

In summary, Experiment 4 focused on the role of bitonal pitch accent alignment in the interpretation of rising and falling pitch contours. Whereas Experiment 3 showed that higher scaling of a bitonal pitch accent may increase the likelihood of an orthogonal interpretation that leads to

[30]The question of how much time is needed for the implementation or perception of plateaus is outside the scope of this work, although Steffman & Cole (2024) shows that the distinction between H-H% and H-L% in production is small but not further enhanced with additional syllables following the accented syllable.

reduced proportion of TELLING responses for falling contours, Experiment 4 has shown that manipulating the alignment of that high-scaled bitonal accent (between L+H* and L*+H exemplars) does not play a large role in further modulating interpretation for falls.

2.8 Including a Third "Other" Option (Exp. 5)

Based on the results of Exp. 1-2c, it was hypothesized that the increased prominence for falls may make focus-enriched interpretations more likely, leading to a response behavior wherein participants used the ASKING response option as an ad-hoc "not-TELLING" option. Although an appeal to focus was motivated by prior work on prominence and prosodic focus in MAE (and other Germanic languages) and later supported by the results of Exp. 3 (using L+H*), whether such focus-enriched interpretations are salient remains somewhat speculative. To be clear on this point, it is uncontroversial to say that high prominence with falls is often reflective of prosodic focus marking on a constituent and used to convey communicative functions like contrast or surprise—**focus** is not speculative in this regard. Rather, what is at issue here is whether these same communicative functions are at play in accounting for the patterns observed in the experiments using the current paradigm.[31] The 2AFC paradigm used in Exp. 1-4 is notably limited: participants are exclusively restricted to provide only ASKING or TELLING responses, making the effect of accentual prominence and the interfering alternative interpretations indirect. What is needed now is to identify whether the focus-enriched interpretations that were appealed to when discussing variation in the results—INCREDULITY, CONFIRMING, LISTING, etc.—are indeed the kinds of interpretations that participants have in mind when making their responses in the task.[32] To do so, the final experiment will extend the 2AFC task to take an exploratory look at participants' interpretations by providing

[31] To reiterate: The pattern of interest is that higher accentual pitch appears associated with a lower, not higher, likelihood of assertive interpretations for falls. When the high accentual pitch target is made more prominent (as in Exp. 3), the likelihood of assertive interpretations falls even more.

[32] One may object on formal grounds that if focus is indeed at play here, then of course it would come as a result of the meanings that focus is known to convey. For instance, it would be unsurprising for contrast to be mentioned but completely unprincipled and unexpected for participants to mention something unrelated such as animacy. However, what is at issue here is whether focus plays a role **in the context of the task**; there are other possibilities that could affect task performance such as specific contours being perceptually ambiguous or participants instead opting to lean on paralinguistic attributes that are more intuitive for them to describe, such as affect or emotion.

a "catch-all" OTHER response, which is later followed up with free-text responses.

2.8.1 Procedure

Participants perform a three-alternative forced choice (ASKING/TELLING/OTHER) task using a subset of 34 contours used in Experiments 2, 3, and 4 (shown with our results in Figure 2.16).[33] As in previous experiments, participants listened to 5 repetitions of each contour, with repetitions spread evenly across 5 blocks of trials and presented in a pseudorandomized order for a total of $34 \times 5 = 170$ trials. After completing the entire 3AFC portion of the experiment, participants were tasked with providing free-text responses about their interpretation for any contours they gave an OTHER response to. Participants were given a screen that paired individual audio players with small text boxes (one per contour) so that they could replay the audio files as many times as they needed. Some steps of the continuum were hard-coded to require participants to give free-text responses, but if participants responded OTHER to contours that were not hard-coded then they gave free-text responses to those too—but no more than 20 free-text responses in total. Additional discussion of the structure of the task is provided in Appendix A.4.

2.8.2 Results

A new group of participants who did not participate in any of the previous experiments was recruited from Prolific (n=45). Three participants were excluded, leaving a total of 42 participants (24F, 18M, average age 43.2) for analysis. The 34 contours presented in this experiment are shown in Figure 2.16 along with the proportion of OTHER (versus ASKING or TELLING) responses and the proportion of TELLING (versus ASKING, for responses that were not OTHER) responses.

The prediction for this experiment is that interpretations other than ASKING or TELLING would be absorbed by the OTHER response, leaving the remaining proportion of TELLING responses at

[33]Each experiment used 25 contours, so this subset of 34 contours is out of a total of 75 contours. The contours were selected to provide a broad coverage of each experiment's range of phonetic variation while still keeping the experiment size manageable for participants; here, each block is 9 trials longer than before. The specific contours were selected to present a variety of contour shapes to the participant while still enabling a comparison between experiments. Contours 4-1 and 4-5 from Experiment 3 were intended to be included, but due to experimenter error contours 2-1 and 2-5 were used instead; this does not affect the current results.

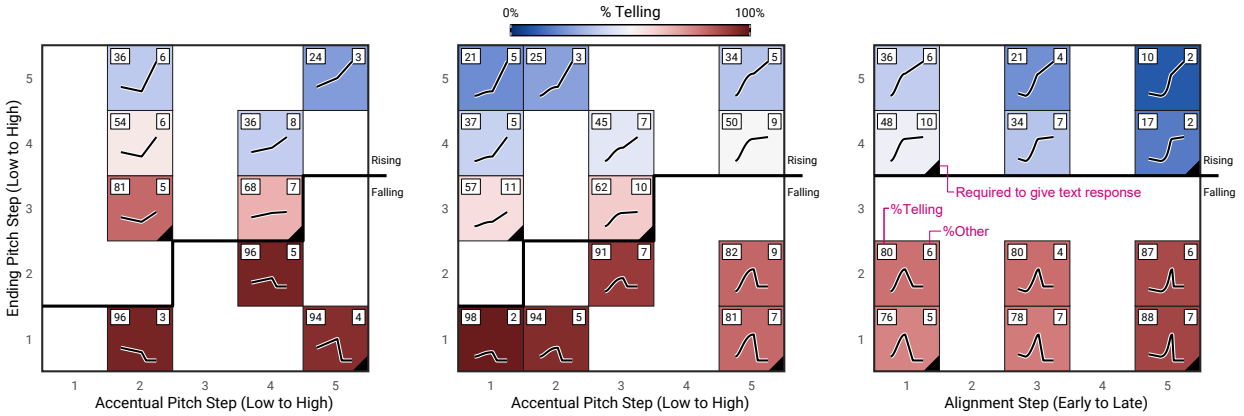


Figure 2.16: Exp. 5 (3AFC with OTHER response) results. The proportion of TELLING responses is shown in the top left of each cell (as in previous heatmaps) while the proportion of OTHER responses is shown in the top right corner. Cells tagged with a black triangle in the lower right corner denote contours participants were required to respond to even if they never respond Other.

floor (for rises) or ceiling (for falls). Based on the heatmaps in Figure 2.16, we can see that such a ceiling/floor effect in the 3AFC task was not obtained. Somewhat surprisingly, participants did **not** frequently use the OTHER response option. Across all contours, the range for the proportions of OTHER responses (vs ASKING/TELLING) for an individual contour was merely 2% to 11% of responses. However, the heatmap display admittedly makes it difficult to identify patterns between the 3AFC task and the previously presented 2AFC tasks; as such, Figure 2.17 compares the results from the previous experiments to the current experiment.

From the results in Figure 2.17 we can see more clearly that the participant response behavior generally did not move closer to the ceiling (for falls) or the floor (for rises). Rather, responses generally shifted towards chance. This is most evident for the rising steps, where the proportion of TELLING responses raised to be closer to chance. The rising steps that showed the opposite pattern (increasing away from chance, i.e., closer to ceiling) were those with very shallow pitch excursions, which again were more like plateaus; these steps were already above 50% in the 2AFC task and generally rose closer to ceiling. Notably, when comparing the proportion of OTHER responses to the proportion of TELLING responses, the proportion of OTHER responses was highest for steps where the proportion of TELLING responses were closest to chance. In other words, while participants generally did not use the OTHER response option, opting instead to accommodate a

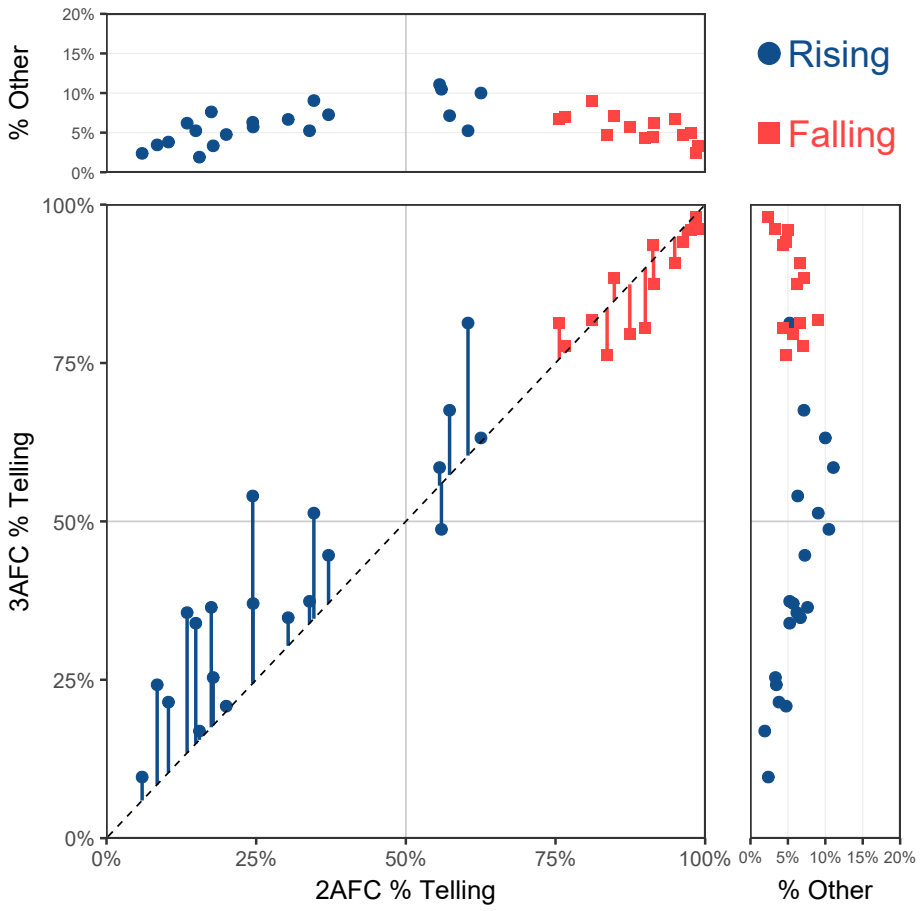


Figure 2.17: Comparison between the 2AFC and 3AFC task results, where the proportion of TELLING responses from both sets of experiments are plotted against one another in the main panel. The smaller panels additionally shows the proportion of OTHER responses from the 3AFC task compared to the proportion of TELLING responses from the 2AFC results (top) and the 3AFC results (right). Note that % TELLING in the 3AFC task is the proportion of TELLING responses out of the total number of TELLING or ASKING responses after setting aside the OTHER responses.

contour as ASKING or TELLING. The contours that were most likely to receive a rare OTHER response were those where participants were at chance in deciding between either ASKING or TELLING.

2.8.3 Free-text Responses

Although participants rarely used the OTHER response option, the setup of the experiment still provided a variety of free-text responses to many contours. Recall that the hypothesis for some of the counterintuitive patterns in the previous experiments was that there existed focus-enriched

interpretations that detracted from a straightforward mapping along the Q/A contrast. For the falls, it was proposed that focus-enriched interpretations (Krifka, 2008; Rooth, 1992) such as CONFIRMING, CONTRASTING, or CORRECTING may be salient to participants. When extended to the rises, other interpretations related to incredulous and confirmative rising declaratives (Goodhue, 2024) may also be salient to participants.

In total, 581 responses were obtained and subsequently coded by myself as belonging to one of eight thematic categories based on a qualitative review of the responses, described below. The number of responses in each category, along with a description of each, is given in Table 2.7. A few illustrative examples of participant examples will be provided with discussion of the dominant themes in the responses.

Category Title	**Count**	**Percent**	**Description of Category**
`Additional Nuance`	114	19.62	Description of ASKING or TELLING plus some nuance such as surprisal, sarcasm, or uncertainty.
`Distinct Function`	84	14.46	Description of a function that is distinct from (but potentially in addition to) ASKING/telling such as confirming or listing.
`Metalinguistic Uncertainty`	49	8.43	Participant conveys uncertainty about how to interpret the contour, such as saying they are uncertain or saying it could be either ASKING or TELLING.
`Asking`	86	14.80	The response is little more than "Asking."
`Telling`	166	28.57	The response is little more than "Telling."
`Other`	2	0.34	The response is little more than "Other."
`Audio Description`	30	5.16	The participant primarily describes the pronunciation or acoustics of the utterance.
`Unusable`	50	8.61	The response is ambiguous as to what the participant intended or merely reiterates the sentence.
Total	581		

Table 2.7: Cross tabulation of response counts by thematic category.

From the distribution of responses, 43.3% of responses explicitly referenced an INQUISITIVE or ASSERTIVE interpretation. That is, some participants treated the free-text response as essentially

the same as the forced choice task they had just completed and provided responses like "telling" or "This sounds like asking." Some participants who responded with OTHER in the forced-choice portion of the task later reported that "Listening to the audio again it sounds like the speaker is actually asking and I don't have an other response" and "i was wrong on this one, sounds like telling." We can view these kinds of responses under a broad lens of accommodation: Participants are willing to accommodate a particular form-function mapping despite other mappings being available. Given that participants just went through 170 trials of ASKING/TELLING/OTHER forced choice responses, the ASKING/TELLING labels were arguably made to be the most salient functions going into the free-text response portion. Indeed, even when looking ahead to the 198 Additional Nuance or Distinct Function responses, 58 of these (29%) describe their interpretations as ASKING/TELLING **plus** something else. While participants were encouraged to consider other interpretations in their responses, it was likely easier to accommodate one of the two labels they had become accustomed to.

2.8.3.1 Additional Nuance Responses

Of the responses coded as Additional Nuance, the top two nuances that participants described were related to *surprise* (≈28% of Additional Nuance responses) and *speaker uncertainty* (≈9%). A few illustrative examples are provided in Table 2.8, where specific contours are denoted by Experiment-Column-Row indices (where Experiment has values of 2, 3, and 4, shown from left to right in Figure 2.16). Looking at the additional nuances, the most frequent responses are related to surprise or speaker uncertainty. While these can occur with either falls or rises, they were most frequent with high-scaled accentual pitch targets. In fact, of the 34 responses noting a nuance of *surprise*, 20 of these were from the bitonal alignment materials, which all had the same accentual peak height. This appears to be in line with a paralinguistic view such that more extreme pitch excursions can convey greater speaker engagement or emotional activation (Goodhue, 2024; Gussenhoven, 2004). So, while incredulous rising declaratives appear to be salient to participants, incredulous falls with high-scaled pitch accents are also salient, suggesting that incredulity as cued by more extreme pitch

excursions may have played a role in some of the counterintuitive patterns previously observed.

Nuance	Contour	Participant Response
Surprise	4-3-1	*asking like they're surprised / exclaiming*
	4-5-4	*The speaker sounds somewhat more suprised then actually asking.* [sic]
	4-3-1	*He is stating it with surprise, but still telling.*
Speaker Uncertainty	3-5-4	*I think he's asking, but is uncertain of his wording*
	3-5-5	*speaker not sure*
	4-5-2	*sound skeptical or uncertain* [sic]

Table 2.8: Verbatim examples of participant responses coded as Additional Nuance.

2.8.3.2 Distinct Function Responses

Looking at the distinct functions, shown in Table 2.9, the most frequent functions that participants noted were *listing* (46/84 responses) and *confirming* (16/84 responses).[34]

Function	Contour	Participant Response
Listing	3-3-3	*He is telling and listing.*
	4-3-4	*The speaker is reciting a list.*
	2-4-3	*Telling as if listing off where people are from.*
Confirming	4-5-4	*like to confirm*
	3-5-5	*telling, but restating as if repeating what he just heard for emphasis*
	4-5-2	*Asking a question back by stating it like they can't believe it and needs someone to confirm.*

Table 2.9: Verbatim examples of participant responses coded as Distinct Function.

The contours receiving listing functions were, unsurprisingly, primarily the contours that had plateau-like shapes, though there were some rises that received listing interpretations as well (see Burdin & Tyler, 2018 for discussion of interpretations of rises and plateaus in lists).[35] For the

[34]For the *confirming* responses, some descriptions of confirmations were in terms of the speaker confirming an answer (i.e., TELLING and CONFIRMING) or the speaker soliciting a confirmation (i.e,. asking to confirm), which are considered together here under a broad category of *confirming*.

[35]Note that while the association between listing and plateau contours has been well described (see Burdin & Tyler, 2018, pp. 98–100 for a review), whether "listing" should be formally understood as a distinct speech act from assertion or whether it is in some sense layered upon assertion is not at all explored in this work. The goals of this work focus on variation in intonational form, not on the semantic or pragmatic characterization of assertions or listing. What is

confirming responses, these were most frequent with falls containing high-scaled pitch accents.

2.8.3.3 Metalinguistic Uncertainty Responses

Finally, the Metalinguistic Uncertainty responses shown in Table 2.10 are a bit more straightforward. Given the previous sets of results, we know that some contours are compatible with particular nuances or functions that are not straightforwardly ASKING or TELLING. Yet, for a participant trying to accommodate such a nuance or distinct function as either ASKING or TELLING, there may be uncertainty as to which they should pick. In this set of responses, some participants did indicate that they themselves were unsure of whether something should be interpreted as ASKING or TELLING. Sometimes this uncertainty was explicit, where the participant explicitly says they don't know or are unsure, or implicit, where the participant says it could be either ASKING or TELLING.

Uncertainty	**Contour**	**Participant Response**
Explicit	3-1-3	*Im not sure this sound like telling* [sic.]
	2-4-3	*not sure, something in between asking and telling*
	2-4-3	*I have no idea on this one, truly.*
Implicit	3-1-3	*Could be either telling or asking so other could be appropriate*
	4-1-4	*The nuance is kind of in the middle.*
	2-2-3	*I think this person is asking, but the tone is close enough that it might be telling, also.*

Table 2.10: Verbatim examples of participant responses coded as Metalinguistic Uncertainty.

Taken together with the Additional Nuance and the Distinct Function responses, the Metalinguistic Uncertainty responses show that while some people are able (or at least willing, for this task) to posit a distinct or nuanced interpretation, others do not do so quite as readily.

relevant here is that participants' responses, which mention listing, reflect prior description that plateaus are used with lists.

2.8.4 Discussion of Free-Text Responses

This section presented a dual 3AFC and free-text response paradigm, with an exploratory goal of identifying whether some of the alternative focus-enriched interpretations proposed in the discussion of the previous experiments were actually salient to participants. Somewhat surprisingly, participants were not particularly eager to use the OTHER response in our task. Some participants were even less eager to elaborate in their free-text responses, suggesting an overall strategy of accommodating the ASKING/TELLING response options. Moreover, the primary focus-enriched interpretations that were considered in §2.5.1 (contrasting, confirming, and mirativity/surprise) showed differing degrees of salience to participants. For instance, 'contrast' was not mentioned by any participant while many participants mentioned speaker surprise.

While the hypothesis regarding focus-enriched interpretations was motivated by increased prominence for the falling steps (resulting in a 'non-TELLING' response strategy), a similar pattern for the rising steps (where lower accentual pitch targets yield higher proportions of TELLING responses) was not seen in the previous experiments. Such a pattern would have been in line with the hypothesis that prominence for low F0 targets is enhanced by lower F0 targets (Liberman & Pierrehumbert, 1984; Pierrehumbert, 1980), but ultimately there was little variation in accentual pitch. Yet, many of the free-text responses commenting on additional nuances and distinct functions were in fact to rising contours. Similarly, many plateau-shaped contours (with very shallow rising pitch excursions) received distinct LISTING interpretations.

Based on these results, it appears that although the canonical associations between intonational form and function (in the narrow context of falling and rising declaratives) are robust given the high and low proportions of TELLING responses for H*L-L% and L*H-H%, respectively, there still remains a task effect wherein other interpretations interfere with probing the Q/A contrast. However, such a task effect does not appear to be exclusively linked to the marking of focus. Setting aside the known association between plateaus and LISTING (Burdin & Tyler, 2018), one potential alternative account may be that participants infer an indirect speech act from the speaker's use of a rising declarative or a fall with high prominence. For instance, one participant described

a fall with high accentual pitch as "Asking a question back by stating it like they can't believe it and needs someone to confirm." Here, the speaker is perceived as indirectly asking a question by virtue of expressing their surprise when stating the information. Similarly, Jeong (2018) describes how rises can also be used for requests or invitations.

In summary, although the general hypothesis that other interpretations interfere with probing the Q/A contrast—hence creating task-specific response behaviors—is supported by the results of the 3AFC task, the more specific hypothesis that these other interpretations are related to focus is less supported. Although it is possible that the participants in the 3AFC task, compared to the participants in the previous 2AFC tasks, were more likely to consider interpretations beyond the Q/A contrast given the presence of an OTHER option, it seems unlikely that the proposal of competing interpretations would exclusively apply to the results of the 3AFC task given that the L+H* continuum used in Exp. 3 did in fact lower the proportion of TELLING responses compared to Exp. 1-2c. With regard to the free-text responses, the majority of interpretations offered up by participants are also in line with prior discussion of variation in intonational function for rising declaratives. In this light, while participants can accommodate particular interpretations that are made salient, such as broad ASKING and TELLING interpretations, the results presented here suggest that certain intonational features, like high accentual pitch scaling, can increasingly make other interpretations more likely. However, due to participants being fairly willing to accommodate ASKING or TELLING interpretations in their free-text responses, this work is limited in its capacity to narrowly link specific contours to specific competing interpretations.

2.9 General Discussion

This chapter addressed the following broad research question: Which region of the nuclear pitch contour contour matters for interpretation along the INQUISITIVE-ASSERTIVE dimension? While prior work often leans on underspecified descriptions of "steepness," this work has investigated variation in rises and falls through a targeted investigation of a series of "fundamental" parameters that could be independently manipulated: starting/accentual pitch, ending pitch, peak/valley

alignment, and onglide/offglide shape. When possible, these fundamental parameters were related to phonological constructs provided by the AM model for MAE. Two competing hypotheses relating these parameters to interpretation along the INQUISITIVE-ASSERTIVE dimension were proposed: an **Edge-only** Hypothesis, where only the edge-tones matter, and an **Integrative** Hypothesis, where both the pitch accent and the edge tones matter for interpretation. On the surface, that there were **any** effects of accentual pitch seems to give support for the Integrative Hypothesis: The choice of pitch accent affects the likelihood of ASSERTIVE interpretations given a rise or fall of particular phonetic expression. However, the effect of accentual pitch was generally not consistent across rising and falling tunes: Higher accentual pitch for falls lowered the proportion of TELLING responses while the effect for rises was more mixed. Table 2.11 provides a summary of the statistical modeling results for each experiment.

	Experiment					
Effect	1	2	2b	2c†	3	4
Accentual F0	$-$	$\times$	$-$	$\times$	$\times$	NA
Ending F0	$-$	$-$	$-$	$-$	$-$	$-$
: Accentual F0	$\times$	$\times$	$\times$	$+$	$+$	NA
Accentual F0 | Fall	$-$	$-$	$-$	$-$	$-$	NA
Accentual F0 | Rise	$-$	$-$	$-$	$+$	$\times$	NA
Alignment	NA	NA	NA	NA	NA	$\times$
: Ending F0	NA	NA	NA	NA	NA	$-$

Table 2.11: Summary of statistical results across experiments. Credible evidence for an effect is indicated by the direction of the effect ($+/-$), effects lacking evidence to suggest they are different from zero are indicated by an $\times$. Effects that are not applicable to an experiment are shown with NA. Experiments: (1) Monotonal accents, (2) Monotonal accents with early falls, (2b) Monotonal accents with early falls and longer durations, (2c) Monotonal accents with early falls and varying alignment, (3) Scaling of L+H*, (4) Alignment between L+H* and L*+H. †Accent alignment covaries with Accentual F0 for Exp. 2c, so only Accentual F0 is considered here.

Work in semantics and pragmatics on rising declaratives has proposed, at minimum, a distinction between steep INQUISITIVE rising declaratives and shallow ASSERTIVE rising declaratives (Jeong, 2018). However, there may yet be subtypes within the ASSERTIVE class of rising declaratives such as INCREDULOUS and CONFIRMATIVE rising declaratives, which have been proposed

to relate to variation in the slope of the rise (Goodhue, 2024). The results of the experiments presented in this work show that the distinction between shallow and steep rises at best relates only somewhat to the choice of pitch accent. In Experiment 2c, the proportion of TELLING responses were slightly higher for rising contours with higher accentual pitch, but even these contours were still overwhelmingly interpreted as INQUISITIVE. The shallow rises of the sort Jeong (2018) and Goodhue (2024) describe are likely best described by variation in the ending F0 target: lower ending F0 leads to markedly less INQUISITIVE interpretations. Moreover, some of the variation in the results appears driven by task-specific effects related to interfering interpretations. Such interference was found to be more robust in Experiment 3, where using a bitonal L+H* decreased the proportion of TELLING responses for falls. Free text responses elicited in Exp. 5 additionally spoke to the availability of interpretations related to surprise, uncertainty, listing, and confirmation.

What do these results suggest for the status of between- and within-category variation in the predicted inventory of tunes from the AM model for MAE? The results from this chapter showed a robust distinction between H-H% and L-L% edge-tones, mirroring findings from production tasks Cole et al. (2023). When looking at the capacity for categorical distinctions based on the pitch accent, the results are more mixed. With the rising contours tested in this work in particular, contours closest to H*H-H% were consistently INQUISITIVE, casting doubt on a robust distinction between H*H-H% and L*H-H% in the context of the Q/A contrast as described by prior work. Rather, variation of the final F0 target (i.e., the acoustic correlate of the edge tones, potentially within a broader rising category) appears to be a more robust cue that delineates between INQUISITIVE and ASSERTIVE rising declaratives. Whether H*H-H% is categorically different from L*H-H% along some other meaning dimension is certainly still possible (e.g., perhaps in terms of surprisal as shown in Dutch by Gussenhoven & Rietveld, 2000), but based on the results of this chapter, they do not appear to be differentiated by the Q/A contrast. The results of this chapter are also compatible with a view that variation in rising tunes may merely reflect meaningful gradient phonetic variation within a single broad category rather than four distinct categories that differ in the pitch accent specification. What the results presented in this chapter show is that deviating from a proto-

typical INQUISITIVE rise—one that starts low and early and ends high—yields comparably fewer INQUISITIVE interpretations. Variation in the ending F0 target heavily influences interpretation along this dimension, but variation elsewhere shows only a limited effect on interpretation beyond perhaps sounding like a less clear exemplar of what an INQUISITIVE rise should sound like.

2.9.1 Limitations and Future Work

This work has focused primarily on testing the potential roles of structured phonetic variation in different regions of rising and falling intonation for interpretation along the Q/A contrast. The materials used throughout the experiments in this work have expanded greatly on the materials used in Jeong (2018), but it is nonetheless worth reiterating that it was **not** a goal of this work to provide a refined or updated semantic/pragmatic analysis of different types of rising declaratives. The connection to Jeong's work is focused entirely on limitations in that work related to the phonological proposal (that the divide between IRDs and ARDs is between L*H-H% and H*H-H%) and the phonetics of the materials used to support such a proposal. However, this work is crucially limited by not incorporating a prior discourse context for participants to use in interpreting our materials. The interpretation of rising intonation has been shown to be highly sensitive to factors beyond the phonetic expression of the pitch contour to the extent that variation in the phonetic expression can be largely overridden by contextual demands, for instance the relative knowledgeability between the speaker and their addressee (Goodhue, 2024; Jeong, 2018). Rudin (2022, pp. 343–344) also discusses four empirical generalizations capturing the felicity of rising declaratives, one of which being "An utterance of $p?$ is only felicitous when the speaker has reason to believe that the addressee believes p." Again, because no context was given, it is possible that some participants found that our decontextualized rising contours were not felicitous, and hence were not able to accommodate an INQUISITIVE interpretation that requires contextual evidence.[36] At the same time, some participants may have been more readily able to construct such a context that would support the INQUISITIVE interpretation. Work on cue combinatorics with regard to prosody and pragmatics

[36]Note however that these same limitations were shared by Jeong's design.

is quite limited (see, e.g., Ronai et al., 2019), and so future work may opt to investigate the nature of participant response behavior under discourse manipulations that, for instance, either do or do not license (or *favor*, in less strict terms) the use of specific intonational tunes. Similarly, one may investigate the role and importance of intonation when it is either made redundant by the context compared to when it is the only cue available (as in the experiments presented in the present work). Practically speaking, such experiments likely would not need the large amount of phonetic variation that was explored here, and so these results can serve as a starting point for choosing which distinctions to explore further.

An additional limitation of this work is that it is restricted to MAE, and so these results are certainly not intended to be representative of all languages or dialects, which may differ in their use and distribution of rising intonational patterns. However, it would be an interesting avenue to explore whether the factors that did not play a large role interpretation are in fact important in other languages or dialects.

Another limitation is that this work only investigated variation in the nuclear interval of the utterance. One could imagine a series of experiments similar to those presented here, where the prenuclear region is manipulated in terms of the type and scaling of the prenuclear accent on *Molly* or the overall pitch level of the prenuclear region (e.g., high, mid, or low). For example, in a study on Northern Standard German, Petrone & Niebuhr (2014) found that variation in the prenuclear region played a role in participant interpretation of the nuclear accent (see also Roessig, 2024 for the relationship between prenuclear and nuclear prominence in German). Alternatively, one could follow work in Castilian Spanish from Face (2007) by using a gating paradigm to identify prenuclear cues to the Q/A contrast. Recent work from Txurruka (2023) has similarly suggested that Spanish L2 learners are sensitive to the prenuclear accent in identifying Spanish declarative versus interrogative sentence type (which is distinguished by intonation alone).

2.10 Conclusions

This chapter presented a series of perception tasks exploring how phonetic variation in rising and falling intonational tunes relates to phonological distinctions in Autosegmental-Metrical theory for MAE based on interpretation of the Q/A contrast. A robust effect of the scaling of the final F0 target was found: High targets corresponding to H-H% edge tones yield INQUISITIVE interpretations while low targets corresponding to L-L% yield ASSERTIVE interpretations. However, the paradigm used in this chapter also revealed counterintuitive patterns (particularly in relation to falling contours) such that higher prominence, as cued through a more prominent pitch accent like L+H* or increasing scaling of the high accentual peak, may lead to an increased likelihood in competing focus-enriched interpretations, which participants may treat as distinct from merely ASKING/TELLING, detracting from their judgments along this dimension. In a followup task including a third OTHER response with additional free-text responses, participants' ASKING/TELLING judgments were closer to chance for rising contours.

These findings were related to prior work on rising declaratives (Goodhue, 2024; Jeong, 2018) to propose that lower scaling of the boundary tone is likely a more robust cue to ASSERTIVE rising declaratives than the paradigmatic choice of an H* pitch accent. More broadly, these results show that the H*H-H% and L*H-H% tunes do not differ with regard to INQUISITIVE versus ASSERTIVE interpretations—though these tunes may nonetheless potentially differ in terms of some other meaning contribution. Future work may elect to investigate the role of the prenuclear region in conveying distinctions related to the Q/A contrast.

Chapter 3

COMPOSITE MEASURES FOR RISES AND FALLS

3.1 Introduction

Chapter 2 presented a series of experiments on rising and falling intonation in the context of the inquisitive/assertive contrast in Mainstream American English (MAE), with a primary focus on unpacking notions of *steepness* often used to describe variation in rising intonation. Rather than describing manipulations strictly in terms of steepness, the analytic approach employed in Chapter 2 instead opted to describe F0 manipulations in terms of constructs made available within Autosegmental-Metrical (AM) theory (Beckman & Pierrehumbert, 1986; Ladd, 2008; Pierrehumbert, 1980). Specifically, the analysis was in terms of accentual pitch and ending pitch as F0 correlates of the pitch accent and boundary tone, respectively. Importantly, the steepness of a rise or fall is a *composite measure* that depends on these accentual and ending pitch values. One lingering question from Chapter 2 is whether such composite measures are predictive of variation in participants' interpretation of rising and falling intonation. Thus, this chapter revisits the experimental data reported in Chapter 2 to offer an analysis in terms of three composite measures: pitch excursion, slope, and tonal center of gravity (TCoG) (Barnes et al., 2012, 2021).

Two observations motivate this investigation. First, recall that the first two experiments in Chapter 2 both manipulated the accentual pitch for monotonal pitch accents (L* and H*) but differed in the trajectory following the accentual peak. Experiment 1 used a gradual linear fall from an accentual F0 peak to the end of the phrase while Experiment 2 had an early fall immediately following the accentual peak and remained low until the end of the phrase. This early fall manipulation resulted in an increased likelihood of assertive interpretations for the falling contours. But, given that the accentual pitch and ending pitch targets were exactly the same in Experiment 2 as with the gradual linear falls in Experiment 1, it is not immediately clear why the early fall manipulation would result in a higher proportion of TELLING responses. If listeners were only tracking variation in F0 for accentual and ending pitch targets, then the results should have been

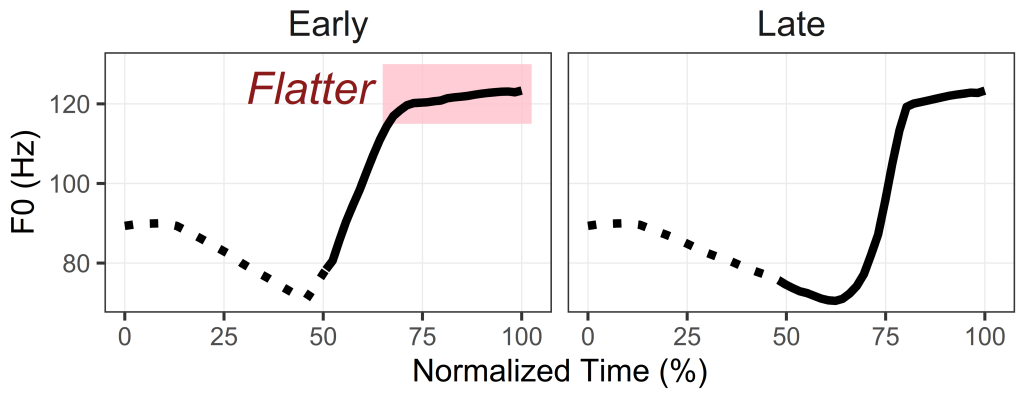


Figure 3.1: Example of the rises with shallow final-rising trajectories from Experiment 4 in Chapter 2. The final rising trajectory is implemented across a greater duration for the earlier-aligned rise (highlighted in pink) than the later-aligned rise, resulting in a flatter rise trajectory.

equivalent. This pattern of results suggests a need to account for something related to the shape of the trajectory beyond merely the variation in these two points in terms of their F0 values—but any such shape distinctions are not accounted for in the models previously presented.

The second observation comes from Experiment 4 in Chapter 2, which manipulated the alignment of the accentual peak and its associated preceding low target between an early-aligned L+H* pitch accent and a later-aligned L*+H pitch accent. In this experiment, variation in alignment primarily affected the shape of rising pitch excursions with very shallow pitch excursions. For these contours, earlier peak alignment resulted in an extended duration for the shallow rise, making them sound more similar to plateaus than rises; examples of these contours are shown in Figure 3.1. This pattern suggests a need to consider the *temporal* duration of the F0 excursion.

The goal of this chapter is to assess the extent to which models based on the composite measures of pitch excursion, slope, and tonal center of gravity are predictive of variation in interpretation along the Q/A dimension in comparison to the models presented in Chapter 2, which used only variation in accentual pitch and ending pitch as predictors. By doing so, it is possible that one or more of these composite measures will offer a potentially more parsimonious account for the observed variation in participants' interpretation of these intonational patterns.

3.1.1 Measures

In this section I describe how the three different composite measure—excursion, slope, and Tonal Center of Gravity—are defined and measured in the context of rising and falling intonation. When necessary, caveats related to the way the stimuli in Chapter 2 were constructed will also be discussed. Recall that the stimuli for experiments in Chapter 2 were declarative sentences like *Molly's from Branning*, where *Branning* is the phrase-final word that receives the nuclear pitch accent. When discussing these measures, we are only concerned with the F0 contour over *Branning*, which was resynthesized to various rising and falling contours. The five sentences comprising the stimuli were resynthesized such that the syllable durations across all sentences were exactly the same, hence the composite measures are the same across sentences.

The first composite measure investigated in this work is the F0 EXCURSION, which measures the direction and total difference between the accentual pitch and ending pitch targets in semitones. Falling contours of the continuum have a negative excursion while rising contours have a positive excursion. Although excursion size is not a measure of steepness specifically, the two measures are nonetheless linearly related. For example, for two contours of the same duration, shallower contours have excursions of smaller magnitude (closer to 0) while steeper excursions have larger magnitudes.

The second composite measure is SLOPE, which measures the change in semitones (i.e., the excursion, Δst) over some time span (Δt)—thus slope is related to excursion by some scaling factor. In the context of the materials in experiments presented in Chapter 2, Δt is the time it takes for F0 to reach the phrase-final target following the accentual pitch target. For the most part, Δt is equivalent to the second syllable duration of the phrase-final word in the sentences (e.g., *ning* in *Branning*). However, in experiments that include an alignment manipulation (Experiments 2c and 4), Δt varies slightly from the same measurement in the other experiments. Additionally, Δt is 70% shorter for early falls than for gradual falls (i.e., the low target is aligned at 30% of the second syllable). Hence, the early falls have steeper slopes than the gradual falls despite the excursion being the same.

The last composite measure considered here is Tonal Center of Gravity (TCoG) (Barnes et al., 2012, 2021; Sostarics & Cole, 2023b), which merits a bit more explanation for the unfamiliar reader. TCoG was originally developed to characterize how onglide trajectories differ for the bitonal L+H* and L*+H pitch accents. Specifically, it has been observed that these pitch accents differ not only in their peak alignment (L+H* is earlier, L*+H is later) but also in the shape of the rising onglide (L+H* is more domed, L*+H is more scooped). TCoG offers a way to characterize these co-varying distinctions in terms of variation in a single, continuous measure. TCoG can be conceptualized as a point in (Time, F0) space describing where the bulk of F0 is. Mathematically, it is the average of time or F0 values weighted by the respective values in the opposite domain (F0 or time, respectively). That is, values contibuting to TCoG in the time domain (TCoG-T) receive higher weights when F0 is higher and values contributing to TCoG in the frequency domain (TCoG-F) receive higher weights when they occur later. This chapter will only investigate variation in TCoG-F; the formula for TCoG-F is shown in Eq. 3.1.

$$\text{TCoG-F} = \frac{\sum_i^n F0_i t_i}{\sum_i^n t_i} \tag{3.1}$$

Figure 3.2 shows how TCoG, denoted as a star, moves between an early domed trajectory and a later scooped trajectory. The key insight here is that the whole trajectory is incorporated into TCoG. The starting and ending F0 values of the trajectories in Figure 3.2 are the same for each trajectory, but TCoG varies because the bulk of each trajectory's F0 movement occurs in different places along the time and frequency dimensions. For the scooped trajectory, despite the high F0 values at the end being weighed more than the earlier low F0 values, TCoG is still somewhat lower than that of the domed trajectory because a greater amount of its F0 is fairly low. In summary, TCoG (specifically in this work, TCoG-F) incorporates information about the entire contour beyond just the pitch excursion or slope.

A schematic depiction of each composite measure is shown in Figure 3.3. What are the predic-

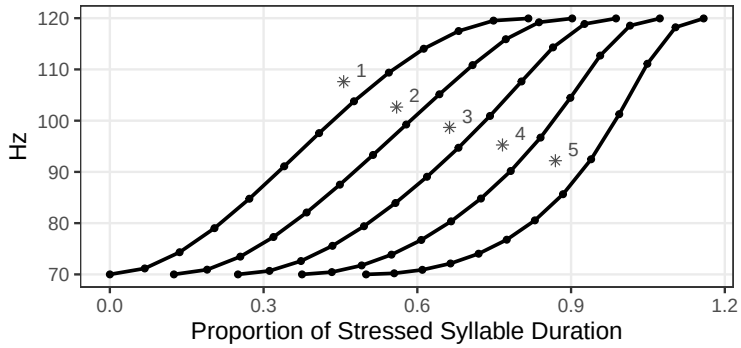


Figure 3.2: Five trajectories between an early domed onglide (1) and a later scooped onglide (5). TCoG is the single (Time, F0) point shown by a star for each trajectory. On the x-axis, 1.0 denotes the end of the stressed syllable.

tions for each of these composite measures? We know from the results of Chapter 2 that variation in ending pitch is a far more robust cue to the Q/A contrast interpretations than accentual pitch is. Because excursion is defined as ending pitch minus accentual pitch (hence varying linearly with ending pitch), larger excursions are expected to have lower proportions of TELLING responses. Slope is predicted to show largely the same pattern, as it is linearly related to excursion. TCoG, being a weighted sum with weights given by the timestamp of F0 samples, will assign the greatest weight to the final F0 sample (i.e., ending pitch), and so we should expect a similar pattern to excursion and slope. In summary, the relationship between each of the composite measures and the results should be very similar to the relationship between ending pitch and the results. More broadly, while the goal of this chapter is to assess how predictive each measure is of the results, we should expect from the outset that each measure will be at least somewhat predictive of the data.

3.2 Methods

To recap, in the experimental task described in Chapter 2, participants were auditorily presented with a declarative sentence such as *Molly's from Branning* uttered with either rising or falling intonation from a wide continuum of contours. Across six experiments, 333 remote participants were recruited from Prolific (mean age 37.8, sd 12.6; 201F, 148M, 12 Other); each experiment had 52 to 56 participants. Participants were tasked with responding whether the speaker was telling them

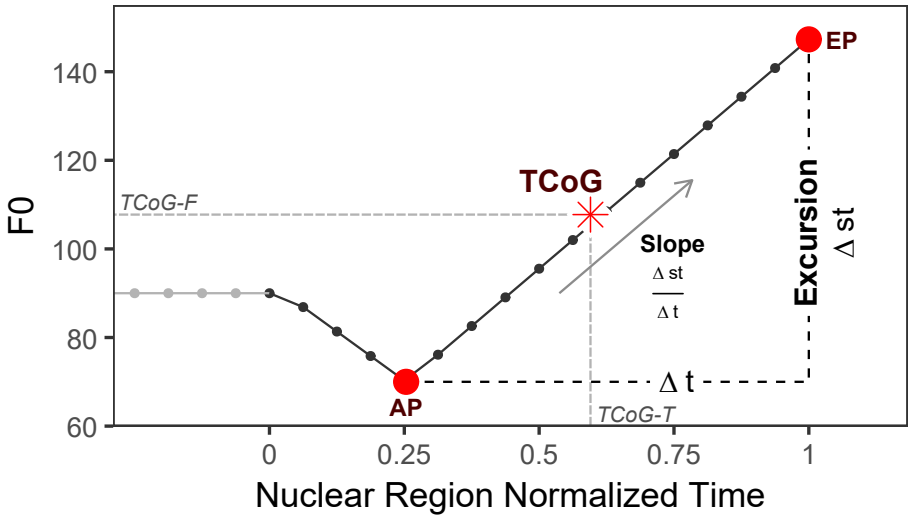


Figure 3.3: Summary of phonetic measures in this work using an L*H-H% contour as an example. The prenuclear region (left, light gray) is truncated for space because it is not used in the measurement/calculation of the different measures. Timestamps used as the weights for TCoG-F are time normalized across the nuclear region, not the entire utterance. AP=Accentual Pitch; EP=Ending Pitch.

something or asking them something, indicating an assertive or inquisitive interpretation, respectively (see also Jeong, 2018 for a similar task). Participants responded to 125 trials in each experiment; each participant had no fewer than 120 observations after removing trials where participants failed to respond within six seconds, leaving a total of 41,451 observations for the analyses in this chapter. The goal here is to see how the proportion of TELLING responses relates to variation in the different measures described in the previous section using Bayesian mixed-effects logistic regression models that use different composite measures as their predictors.

To assess the performance of the different models, the datasets from the experiments presented in Chapter 2 are combined into a single dataset. Models are then fit to this large dataset using 5-fold cross validation.[1] This procedure takes the dataset and creates five "folds" containing a different

[1] An alternative would be to train a model on the results of one experiment and determine how well the model can generalize to the results of a different experiment. However, there are three drawbacks to this approach. First, this approach runs into issues when considering the results of the alignment experiment, as it does not vary in accentual pitch (similarly, the other experiments do not vary in accent alignment) and so would have to be omitted from the current analysis. Second, the model fit to one experiment would be rather fine-tuned to the stimuli of that experiment; for instance, the effect of ending pitch may be of a different magnitude compared to the experiment used as the test dataset, and so would systematically underperform for contours who received more varied responses. Lastly, systematically training n models on one of five experiments, then testing on the other four, is combinatorically cumbersome,

80-20 split of the data: 80% of the data is used to train the model, which is tested on the held-out 20%. This 80-20 split is done at the participant level such that 20% of each participant's data is held out so that each participant is evenly represented in each model.

Model performance on this task is evaluated by a descriptive performance metric and an information criterion metric. The descriptive metric is the area under the curve (AUC) of the receiver operating characteristic curve (ROC curve), which reflects the average trade off between the true positive rate and the false positive rate. For the information criterion metric, the `loo` R package (Vehtari et al., 2017, 2024) is used to compute the 5-fold expected log pointwise predictive density (ELPD) for each model. This measure can be used to compare two models by taking the difference between the ELPD for the two models. When the ELPD difference is small, the models can be considered to make similar predictions to one another.

The models under investigation are described in Table 3.1. A listing of R-syntax formulas are provided in Appendix B.1.

Name	**Description of Predictors**
Scaling	Accentual Pitch, Ending Pitch, and their interaction
Scaling+AL	The Scaling model with an additional alignment term
Excursion	F0 difference (in st) between ending and accentual pitch
Slope	Slope (in st/cs) of the rise or fall from the accent target
TCoG	TCoG-F (in st from 90Hz) of the nuclear pitch contour

Table 3.1: Initial models to use for comparisons.

and it would be difficult to assess overall model performance when the pattern of errors differs by experiment.

3.3 Results

The model predictions compared to the raw data[2] for each model are shown in Figure 3.4. Note that these predictions incorporate the variation from the random effects structure; for a version of the figure in which the participant- and sentence-level variation is "subtracted out," see Figure B1 in the Appendix.

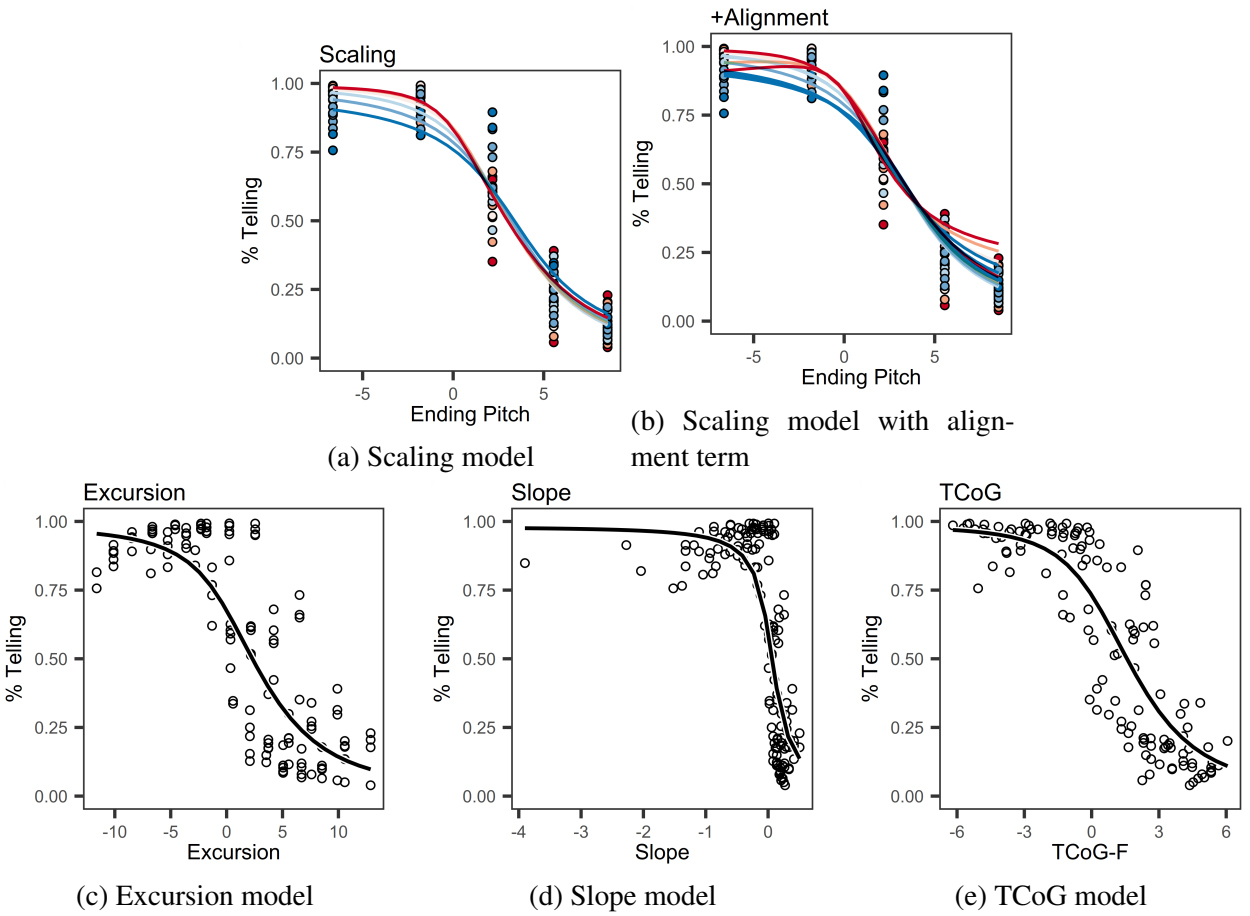
From Figure 3.4, we can see that each model is generally successful at capturing the main patterns of variation in interpretation. The Scaling model shows credible effects of ending pitch ($\hat{\beta} = -0.65, CrI = [-0.69, -0.61]$), accentual pitch ($\hat{\beta} = -0.04, CrI = [-0.08, -0.01]$), and their interaction ($\hat{\beta} = 0.03, CrI = [0.02, 0.03]$). The Scaling+AL model does not show any additional credible effect of alignment ($\hat{\beta} = 0, CrI = [-0.07, 0.07]$) nor its interaction with ending pitch ($\hat{\beta} = 0.01, CrI = [-0.02, 0.03]$). The composite models show credible effects of excursion ($\hat{\beta} = -0.39, CrI = [-0.41, -0.36]$), slope ($\hat{\beta} = -5.71, CrI = [-6.18, -5.24]$), and TCoG ($\hat{\beta} = -0.76, CrI = [-0.8, -0.72]$). Full model summary tables are available in Appendix B.2. These results reflect the general patterns that higher boundary tone scaling is more inquisitive;[3] larger rising pitch excursions are more inquisitive; steeper slopes are more inquisitive; and higher TCoG-F is more inquisitive.

The model performance metrics—the AUC and ELPD sum[4]—are shown in Figure 3.5. A table listing the numeric values is available in Table B10 in Appendix B.2. Also included in Figure 3.5 is a baseline using a model that includes only ending pitch (EP) as a predictor (c.f. the scaling model, which additionally includes accentual pitch (AP) as a predictor). From Figure 3.5 we can

[2]Note that *raw data* here refers to the empirical proportion of TELLING responses for each contour from each experiment. So, contours that appear in multiple experiments appear as multiple data points, i.e., they are not aggregated across experiments. TCoG values are computed directly from the acoustic signal for each file, so the exact values are sensitive to measurement noise which varies on a file-by-file basis, hence the TCoG values for the same contour vary slightly for the different sentences (c.f. the accentual and ending pitch values, which are known constants from the continua). While the actual (file-specific) TCoG values are used in the model, the figure shows the aggregate TCoG values across sentences for each contour.

[3]As in Chapter 2, "more/less inquisitive" should be taken to mean 'is associated with a higher/lower likelihood of inquisitive interpretations' and not taken to be a claim about theoretical "degrees" of inquisitiveness. The same applies, *mutatis mutandis*, for "more/less assertive."

[4]Note that the units of the ELPD sum values are not directly interpretable and the magnitude depends on the number of observations. Here, there are approximately 41,000 observations, each of which has a pointwise ELPD value between 0 and ≈-10, hence the large sum values. See Vehtari et al. (2017) for more information.



(a) Scaling model

(b) Scaling model with alignment term

(c) Excursion model

(d) Slope model

(e) TCoG model

Figure 3.4: Model predictions (curves) versus empirical data (points), where one point equals the average proportion of TELLING responses for one contour from each experiment (total=150). For the Scaling and Scaling+AL models (a and b), the full range of accentual pitch values are shown in color from low (red) to high (blue); only two alignment values (the extrema, 30% and 115%) are used for the Scaling+AL model predictions, shown as separate lines, due to the limited amount of variability. Note that the other three models do not include this information, and so they are not similarly colored.

observe that the Scaling model, when including the alignment term, is the best performing model. However, it should be noted that the performance of the Scaling+AL model is not subsantially higher than the bare Scaling model (which lacks the alignment term); in other words, the addition of alignment in the model offers only a meager performance advantage. Here, the variation in alignment is coming from only two (out of the six) experiments, and even then the variation in alignment is far more limited in comparison to the range of variation in accentual and ending pitch values. Based on these metrics and the previously reported lack of an effect of alignment, it is

not very surprising that the two scaling models behave similarly to each other. Notably, the slope model shows the lowest performance despite including additional information beyond merely the F0 excursion size. However, as is perhaps evident from Figure 3.4, there is an outlier[5] in the slope measurements which likely accounts for the lower predictive performance for the slope model compared to the excursion model.

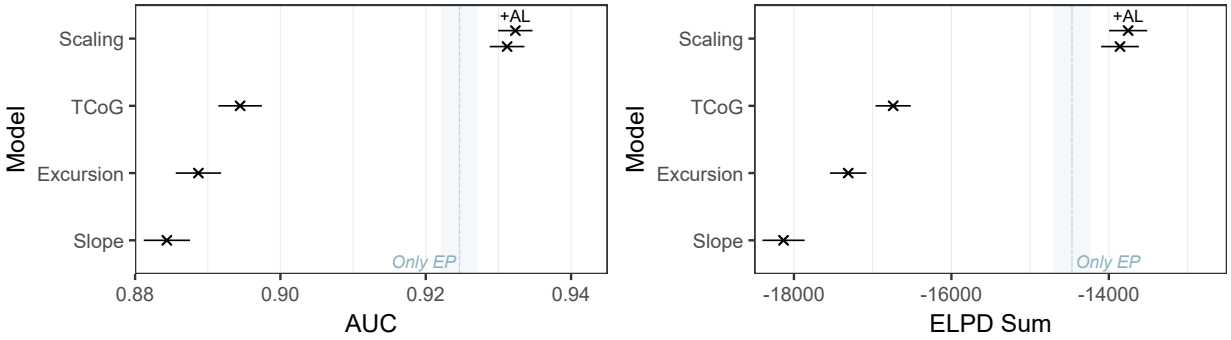


Figure 3.5: AUC (left) and ELPD differences (right) for each model. ELPD differences are are shown as ratios relative to the Scaling+AL model, which shows the best performance. Error bars reflect 95% confidence intervals. The baseline model with only ending pitch is shown by the vertical line labeled "Only EP."

3.4 Interim Discussion

Thus far we have compared a model that includes information about accentual pitch and ending pitch (and an additional model which also accounts for alignment) to three composite models using only one predictor (either excursion size, slope, or TCoG). Relative to a baseline using solely ending pitch, the scaling models perform slightly better. This finding mainly recapitulates what was seen in Chapter 2: Ending pitch accounts for most of the variation in participants' interpretations and adding accentual pitch accounts for a small additional amount of variation. Each of the three composite models perform worse than the scaling model—and worse than a model using only ending pitch. However, it is nonetheless crucial to note that even these "worse models" still perform quite well; for instance, the AUC only ranges from about 0.88 to 0.93 (where at-chance performance would be 0.5).

[5]The specific problematic contour is the latest-aligned L*+HL-L% contour from the bitonal alignment experiment. Here, the late alignment causes the drop in F0 to the early L- target to be very abrupt, resulting in a very steep slope.

While the more parsimonious composite models are predictive of variation in participant responses, one might wonder what kind of information might be missing from these models. For example, take the excursion model predictions shown in Figure 3.5c; while there is a characteristic S-shape in the data expected from a successful two-alternative forced choice task, there remains some variation around this curve that is unexplained. In the context of the materials, one shortcoming of this model is that knowing only the pitch excursion between two points does not give any information about what those points are. For instance, a small rise from a high point is treated the same as a small rise from a low point. Similar limitations can be identified for the other composite models as well. Such limitations suggest that what may be missing from these models is some way to provide structure to the phonetic variation present in the data.

3.5 Adding in Structured Phonetic Variation

This section provides additional model comparisons to evaluate how the performance of the composite models may improve when the models are augmented with additional information about structured phonetic variation. One way to provide structure would be to test whether the magnitude of the effect of each measure differs depending on the global **shape** of the contour (either rising or falling). For example, the likelihood of an assertive interpretation may vary substantially across rises of different magnitudes, but falls may not be as sensitive to such variation in steepness. Alternatively, in line with the previous observation that a shallow excursion may begin from a high or low starting point, we might find that the effect of each measure, in fact, interacts with accentual pitch. Similarly, it may also be the case that the effect of each measure interacts with ending pitch.

The models presented in this section add in either contour shape, accentual pitch, or ending pitch to the existing composite models. In implementing these models, each new predictor is included with its interaction with the composite measure (e.g., excursion size plus shape **and their interaction**). Shape is treated as a categorical variable based on whether the pitch excursion is positive (=rising) or negative (=falling). Accentual pitch and ending pitch are the same continuous variables on the semitone scale used in the scaling model. In terms of notation, the models will

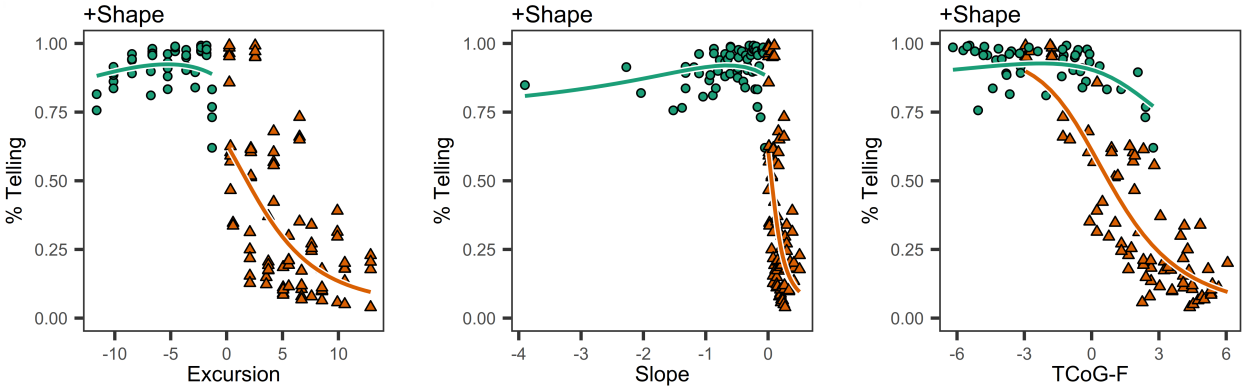
be referred to by appending +Shape, +AP (accentual pitch), or +EP (ending pitch) to the model names (e.g., Excursion+Shape). The model comparison metrics are the same as before.

3.5.1 Results

The +Shape, +AP, and +EP models will be presented in sequence. Note that the addition of new predictors to these models **is not cumulative**. That is, the +AP model does **not** include the shape predictor and the +EP model does **not** include the shape and AP predictors.

3.5.1.1 +Shape Models

All Composite+Shape models show credible effects of their respective measures, rising/falling shape, and their interaction.[6] Figure 3.6 shows the updated model predictions for each composite measure when augmented with a shape predictor. Note that the curves are restricted to the relevant range of the data.[7] For example, the hypothetical model predictions for a falling contour with a positive pitch excursion is not shown.



(a) Excursion+Shape model (b) Slope+Shape model (c) TCoG+Shape model

Figure 3.6: Model predictions (curves) versus empirical data (points), where one point equals the average proportion of Telling responses for one contour from each experiment. Falling contours are shown with green circles while rising contours are shown with orange triangles.

Figure 3.7 shows the model performance metrics updated with the Composite+Shape models.

[6] Refer to the tables in Appendix B.2 for a listing of coefficient estimates.

[7] In the context of the excursion and slope models, including the shape predictor essentially introduces a discontinuity at 0.

We can observe that the inclusion of the shape parameter improves model performance for each of the composite models. Notably, while the slope model originally performed worse than the excursion model (again, due to the influence of the late-aligned bitonal falls) the inclusion of the shape parameter offers more flexibility for the model to handle these points. As a result, the slope and excursion models are more evenly matched.[8] The TCoG model notably performs much better than either the Excursion or Slope models, though it is still not on par with the scaling models.

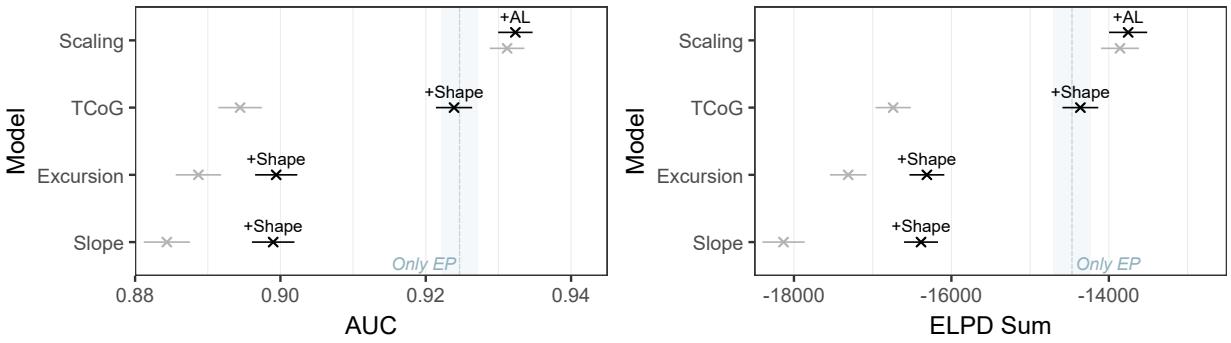


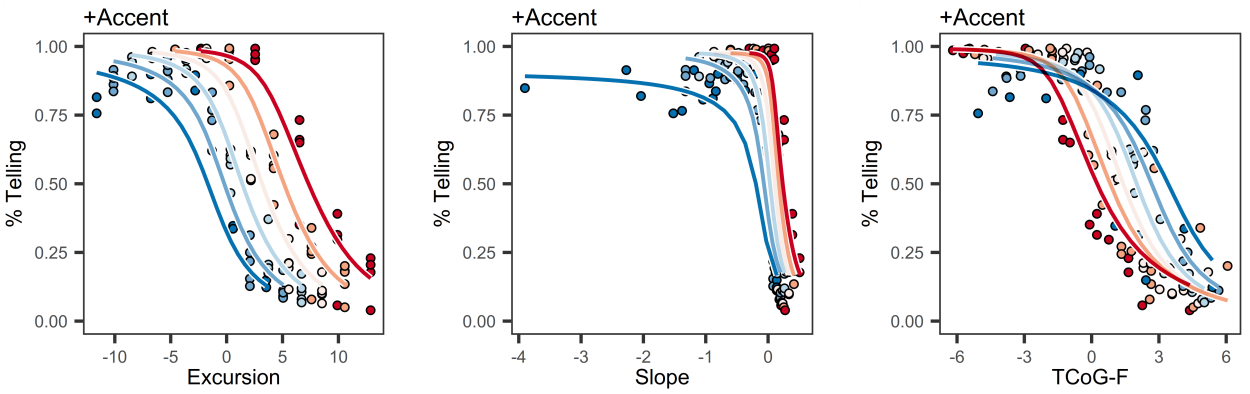
Figure 3.7: AUC (left) and ELPD differences (right) for each Composite+Shape model. Error bars reflect 95% confidence intervals.

3.5.1.2 +Accentual Pitch Models

Each of the Composite+AP models shows credible effects of their respective composite measures, accentual pitch, and their interaction; see the tables in Appendix B.2 for a listing of specific estimates. Figure 3.8 shows the updated model predictions for each composite measure when augmented with a predictor for accentual pitch. We can observe that the addition of accentual pitch affords each model additional flexibility to capture variation in the results.

Figure 3.9 shows the model performance metrics updated with the Composite+AP models. We can observe that including a predictor for accentual pitch greatly improves model performance for each of the composite models to be on par with, or better than, the baseline model using only ending pitch. The Excursion+AP model specifically performs on par with the scaling model.

[8] Again though, given that most of the experiments reported here do not include an alignment manipulation, the excursion and slope models are more or less linearly related to one another, so similar performance after accounting for influential points is not surprising.



(a) Excursion+AP model (b) Slope+AP model (c) TCoG+AP model

Figure 3.8: Model predictions (curves) versus empirical data (points), where one point equals the average proportion of TELLING responses for one contour from each experiment. The full range of accentual pitch values are shown in color from low (red) to high (blue).

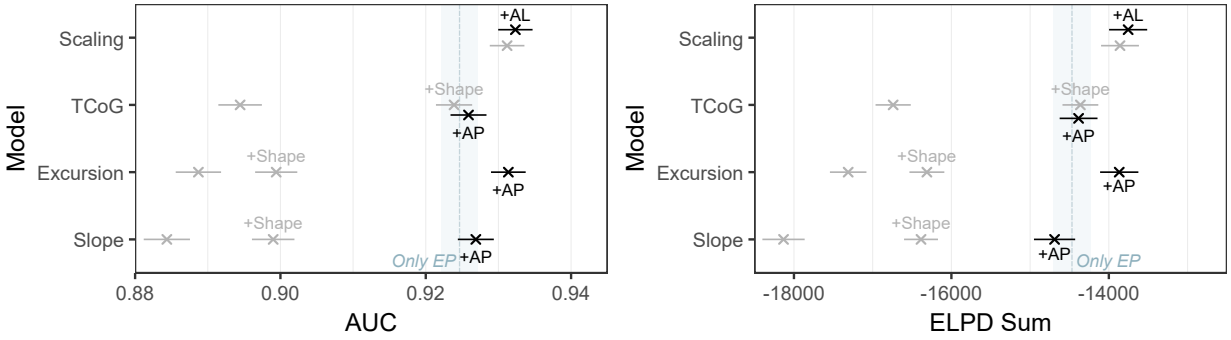
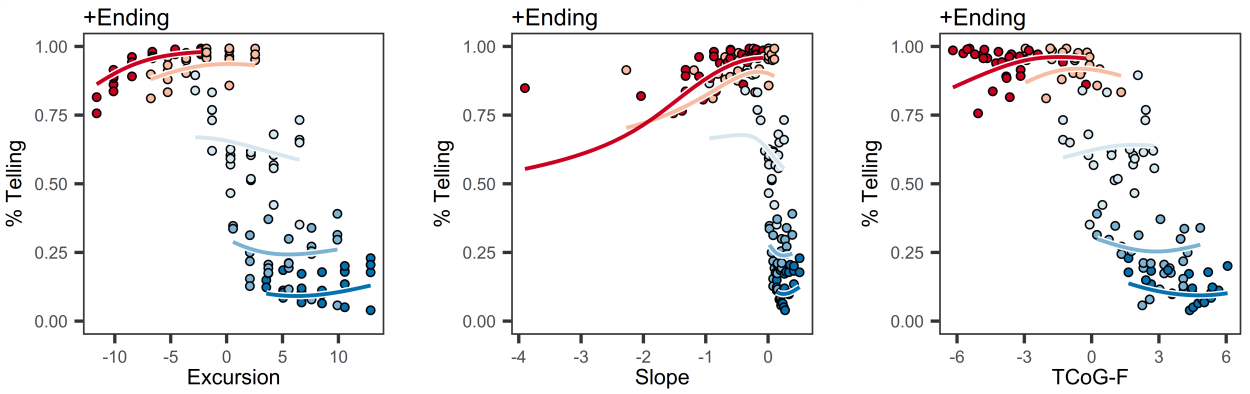


Figure 3.9: AUC (left) and ELPD differences (right) for each Composite+AP model. Error bars reflect 95% confidence intervals.

3.5.1.3 +Ending Pitch Models

As with the Composite+AP models, adding ending pitch to each model provides additional flexibility in capturing variation in the data. Figure 3.10 shows the updated model predictions for each composite measure when augmented with a predictor for ending pitch.

When including an effect of ending pitch, the effect of excursion in the Excursion model ($\hat{\beta} = -0.39, CrI = [-0.41, -0.36]$) was eliminated entirely ($\hat{\beta} = -0.01, CrI = [-0.04, 0.02]$). Similarly, the effect of slope in the Slope model ($\hat{\beta} = -5.71, CrI = [-6.18, -5.24]$) was reduced by an order of magnitude ($\hat{\beta} = -0.72, CrI = [-1.15, -0.3]$) and the effect of TCoG in the TCoG model ($\hat{\beta} = -0.76, CrI = [-0.8, -0.72]$) was over six times smaller and opposite in sign



(a) Excursion+EP model (b) Slope+EP model (c) TCoG+EP model

Figure 3.10: Model predictions (curves) versus empirical data (points), where one point equals the average proportion of Telling responses for one contour from each experiment. The full range of ending pitch values are shown in color from low (red) to high (blue).

($\hat{\beta} = 0.12, CrI = [0.04, 0.2]$). When looking at the model predictions in Figure 3.10 all of these results are reflected by how there is relatively little change in the proportion of Telling responses as each composite measure changes, but broad changes in proportions vary by the groupings of the ending pitch values.[9]

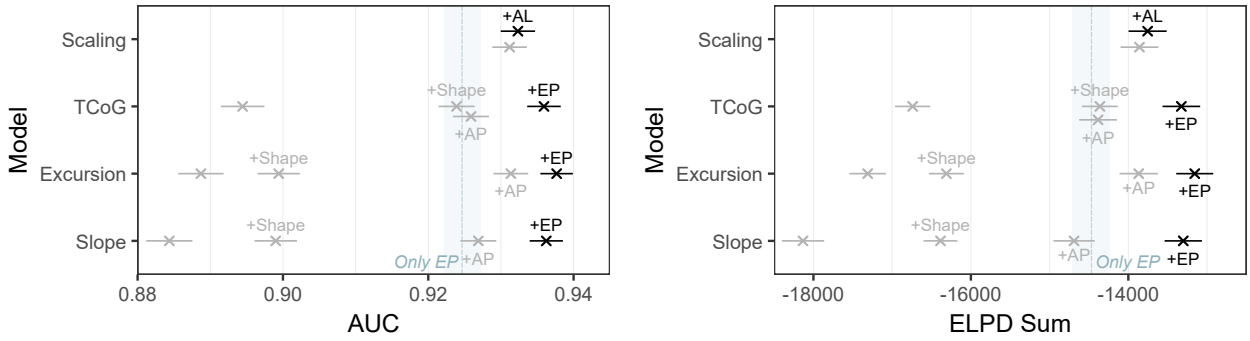


Figure 3.11: AUC (left) and ELPD differences (right) for each Composite+EP model. Error bars reflect 95% confidence intervals.

Figure 3.11 shows the model performance metrics updated with the Composite+EP models. We can observe that including a predictor for ending pitch greatly improves model performance for each of the composite models. Not only that, but each of the Composite+EP models perform better than the two scaling models previously presented. All three of the Composite+EP models perform

[9]This pattern is even more evident when looking at model predictions without including variation from random effects (see the bottom panels in Figure B1 in Appendix B.3).

similarly to one another, although the Excursion+EP model is numerically the most performant model. But, because this model shows **no** credible effect of excursion after including ending pitch as a predictor—yet it appears to be the best model—it is worth briefly considering why adding ending pitch to the models leads to such a substantial boost in performance.

3.5.2 What does Adding Ending Pitch Actually do?

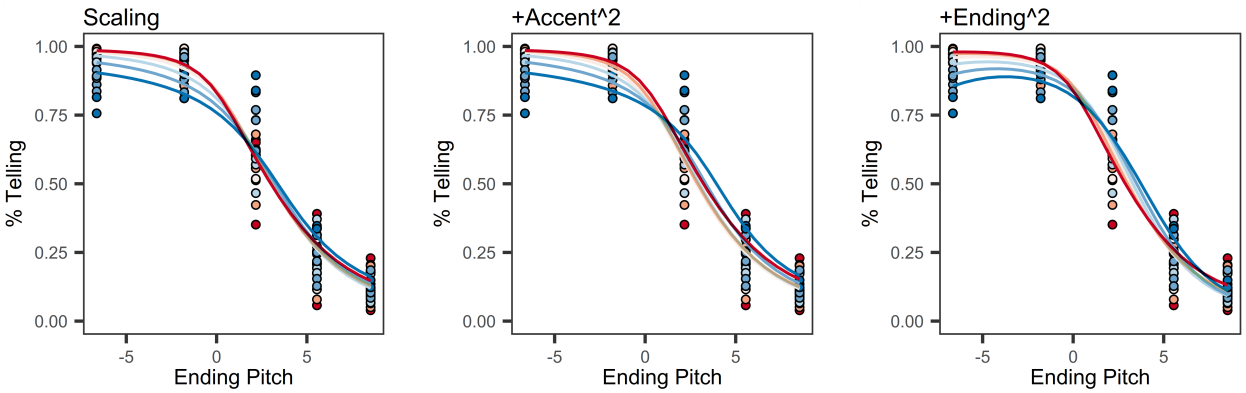
Based on the results of Chapter 2, we know that ending pitch accounts for much of the variation in participants' interpretations. Despite the robust linear relationship between the two, one way to improve the model's performance would be to allow the relationship between ending pitch and the likelihood of a TELLING response to be non-linear. One way to do so would be to add polynomial terms to the model, e.g., adding a quadratic effect of ending pitch; this would afford greater flexibility to the model to capture variation in the data and improve model performance. Although adding a quadratic term (or, optimistically, further including even higher polynomial terms) adds additional flexibility, its inclusion should be theoretically motivated (e.g., see the analysis in Gussenhoven & van de Ven, 2020) to avoid overfitting to spurious relationships or noise in the data. With this point in mind, we now turn to the Composite+EP models.

Why do the Composite+EP models perform better than the initial scaling models? If we consider (for expository ease) the Excursion+EP model, this model contains an interaction term between excursion and ending pitch—mathematically, we get a term of excursion multiplied by ending pitch. But because excursion is defined as ending pitch minus accentual pitch, this means the interaction term can be rewritten as containing a quadratic effect of ending pitch.[10] So, our Excursion+EP model inadvertently introduces a quadratic effect of ending pitch on top of the linear effect of ending pitch that's already included in the model. The insight here is that the Composite+EP models perform better than the scaling models because these models are able to better exploit end-

[10]Short prose derivations showing how the composite models can be rewritten in terms of ending and accentual pitch are provided in Appendix B.1.1. The relevant point here though is that excursion, slope, and TCoG can all be rewritten to contain a $+x_{EP}$ term. When this is multiplied by another x_{EP} term in the interaction, we obtain a x_{EP}^2 term. See also Kock & Gaskins (2016, p. 9) for similar derivations and discussions of quadratic relationships in the context of Simpson's paradox, which will be revisited further in the discussion.

ing pitch, via this extra quadratic relationship, than the scaling models can. We can verify this by adding a quadratic term to the scaling models to close the gap.

Figure 3.12 again shows the predictions from the Scaling model as well as new augmented Scaling models. The Scaling+AP^2 model adds a quadratic effect of accentual pitch while the Scaling+EP^2 model adds a quadratic effect of ending pitch. Qualitatively, we can observe that the added flexbility in the Scaling+EP^2 model accounts for variation in the low ending pitch values (on the left hand side of the figure) compared to the Scaling model.



(a) Scaling model (b) Scaling+AP^2 model (c) Scaling+EP^2 model

Figure 3.12: Model predictions (curves) versus empirical data (points), where one point equals the average proportion of Telling responses for one contour from each experiment. The full range of ending pitch values are shown in color from low (red) to high (blue).

Figure 3.13 shows the model performance metrics updated with the augmented scaling models. We can observe that the Scaling+EP^2, performs on par with the Composite+EP models. Numerically, the Scaling+EP^2 model is also the best-performing model, but its performance is not significantly better than the Excursion+EP model.

3.6 Discussion

While Chapter 2 focused on variation in accentual pitch and ending pitch (recreated here as the Scaling model), there remained a lingering question: to what extent can these results be explained more narrowly in terms of the overall pitch excursion, slope, or TCoG? An account relating interpretation to a single acoustic value would be more parsimonious and would relate more directly

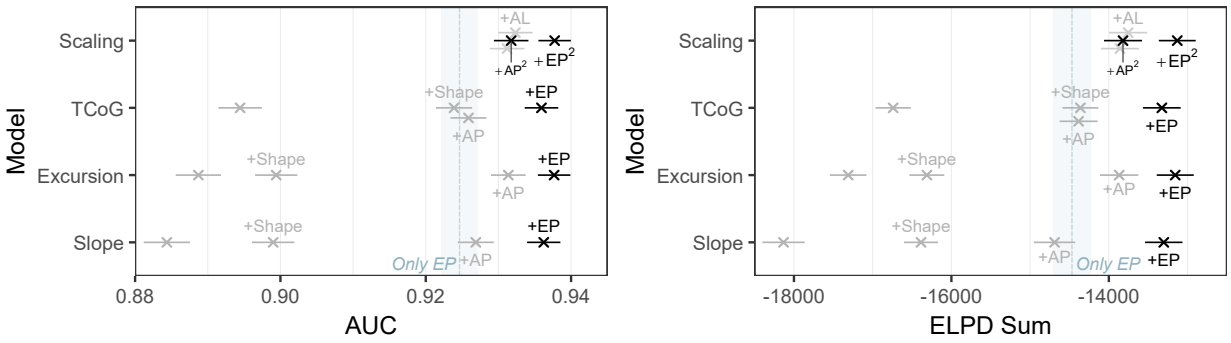


Figure 3.13: AUC (left) and ELPD differences (right) for each model. Error bars reflect 95% confidence intervals.

to coarse-grained observations from the literature about "shallow" and "steep" rises (i.a. Goodhue, 2024; Jeong, 2018). Unsurprisingly, such composite models are indeed predictive of variation in participants' interpretations, suggesting that such cues are relevant to participants' perception. However, models based upon these acoustic measures alone do not perform as well as the scaling models that incorporate information about variation in accentual pitch and ending pitch. Thus, the goal was to identify a relationship where the effect of each composite measure is mediated by another variable describing the phonetic variation.

Additional predictors of rising or falling shape, accentual pitch, or ending pitch were added to the composite models to provide the models with information about how the phonetic variation is structured. While there was a boost in model performance when considering the rising or falling shape of the contour, model performance was much better when considering either accentual or ending pitch. Moreover, adding ending pitch resulted in each composite model achieving better performance than the (initially superior) scaling models. However, adding these terms introduced some complications into the interpretation of the models due to the way the composite measures were related to accentual pitch and ending pitch.

Based on the modeling results showing that the Composite+EP models performed the best, it may be initially tempting to interpret the results as reflecting that participants attend to, e.g., excursion as well as the final F0 target—that is, how much F0 rises or falls, and then where it ends. Yet, there was actually little to no effect of each composite measure beyond what was already

explained by ending pitch. This result seems a bit contradictory: The best model includes both excursion **and** ending pitch, yet excursion itself does not seem to contribute much to the model predictions. This result suggests that ending pitch is a confounding variable such that the apparent effect of excursion is entirely explained by variation in ending pitch.[11] The boost in performance for the Composite+EP models was shown to be related to the implicit quadratic dependency on ending pitch that these models introduced; adding a quadratic term to the scaling model closed the gap between the scaling and composite models. Thus, these seemingly contradictory results when interpreting the Composite+EP models are really more of an artifact of the modeling process rather than a psychoacoustic oddity.

While this discussion seems to invalidate the Composite+EP models, one might wonder whether the Composite+AP results may nonetheless offer a plausible alternative model of the results. After all, when focusing on the Excursion+AP model, there were credible effects of excursion ($\hat{\beta} = -0.63, CrI = [-0.67, -0.59]$) and accentual pitch ($\hat{\beta} = -0.6, CrI = [-0.64, -0.57]$)—and even at similar magnitudes! Again though, the insight that the effect of excursion is confounded by ending pitch is also applicable here, as the Excursion+AP model can be rewritten to contain the terms $(\beta_{AP} - \beta_{Exc})x_{AP} + \beta_{Exc}x_{EP}$. Because the estimates of the effects of accentual pitch (β_{AP}) and excursion (β_{Exc}) were nearly equivalent (≈-0.60), the total magnitude of the effect of accentual pitch is actually quite small ($\beta_{AP} - \beta_{Exc} \approx 0$) and the apparent relationship with excursion is in fact explaining variation related to ending pitch. Although the expository focus has been on the excursion models, these relationships can be similarly shown for the Slope and TCoG models. In summary, although it appeared that the composite models lacked some notion of structured phonetic variation to account for the lower predictive performance, attempts to correct this in practice served to reintroduce ending pitch—and ending pitch itself accounts for most of the variation.

[11] This situation is an example of *Simpson's paradox* (Blyth, 1972; Simpson, 1951), where an apparent relationship between two variables goes away or reverses when controlling for another variable.

3.6.1 Limitations

The model comparisons presented in this chapter were exploratory post-hoc analyses based on the results of Chapter 2. Recall that the goal of Chapter 2 was to use rise/fall "steepness" as a guiding concept to identify regions of the contour where phonetic variation could be introduced. Thus, the experiments tackled variation in accentual pitch (i.e., the F0 correlate of the pitch accent) and the ending pitch (the F0 correlate of the boundary tone). The materials were designed to explicitly vary these dimensions while keeping other factors, such as syllable duration (variation in which would also influence steepness), controlled. The materials were **not** designed to disentangle, say, whether pitch excursion or slope were the primary acoustic cue responsible for interpretation. Similarly, it may be the case that the syllable durations of the nuclear-accented word may also differ in the context of rising versus falling intonation, causing syllable duration to co-vary with slope. While the exploratory comparisons made here seem to suggest that the scaling model vastly outperforms the models using the various composite measures, this finding is not entirely surprising given that the experiment and materials were designed with this model in mind.

These results should not be taken as definitive evidence for or against one composite measure over another. For instance, although the slope model was the worst model tested, neither the materials nor experiments sought to exploit variation in slope specifically. Accordingly, it would be a bit premature to take these results as evidence that listeners do not pay attention to slope. One could imagine (as an avenue for future work) an experiment narrowly looking to disentangle excursion size from slope; for example, where the pitch excursions for a small number of rises are held constant but expressed over longer durations via multiple syllables or variation in speech rate. We could speculate that such an experiment could find that a model equivalent to the slope model presented here is in fact the best performing model for that experiment's results—the opposite of what is found in this chapter. But, crucially, the aims of such an experiment would be very different from the work presented here.

3.7 Conclusions

This chapter has presented a series of model comparisons to explore the extent to which variation in interpretation along the inquisitive/assertive contrast for falling and rising intonation can be explained by different phonetic measures. Specifically, this chapter investigated three composite measures derived from accentual and ending pitch: pitch excursion, slope, and tonal center of gravity (TCoG). It was shown that statistical models including solely these phonetic cues are indeed predictive of variation, but that a model based on variation in accentual and ending pitch performs better. When adding either accentual or ending pitch to the composite models, the model performance (as measured by AUC and ELPD differences) improves considerably. However, much of the improvement in model performance was in fact driven by ending pitch, which was already shown in Chapter 2 to be a robust predictor of variation participants' interpretations.

The analyses presented in this chapter were post-hoc and exploratory in nature. Notably, the analyses were limited in part due to the initial care taken in Chapter 2 to control for potential confounds related to syllable duration and variation in F0. The results of these analyses should **not** be taken as evidence that one composite cue is "the best" nor even better or worse than another cue when considering variation in inquisitive/assertive interpretation. Despite these limitations and the exploratory nature of these model comparisons, the results presented here nonetheless show that ending pitch is a robust cue to participants' interpretation. Scrutiny of the models based on composite phonetic measures repeatedly yielded effects that could be attributed more directly to ending pitch specifically. Future work seeking to more narrowly disentangle excursion from slope would benefit from using materials that were specifically designed to tackle this question.

Chapter 4

THE INTERPRETATION AND PROCESSING OF RISE-FALL-RISE BY WAY OF SCALAR INFERENCE

4.1 Introduction

While it is uncontroversial that intonation contributes to distinctions in pragmatic meaning (Arvaniti et al., 2024; Büring, 2016; Hirschberg, 2017; Ladd, 2008; Pierrehumbert & Hirschberg, 1990; Westera, 2017), what is more controversial is where to draw the line between linguistically contrastive intonational features on the one hand and variation related to meaningful, potentially paralinguistic, gradience on the other (Bolinger, 1978; Gussenhoven, 2004; Ladd, 2008; Ladd & Morton, 1997) The present chapter investigates one tune in MAE that has received ample attention yet nonetheless shows a lack of consensus in terms of describing both its intonational form and function: the RISE-FALL-RISE (RFR) tune. RFR is exemplified schematically in (2), where *italics* denotes the nuclear accented syllable and '...' denotes a phrase-final rise.

(2) A: Did Jane eat all of the cookies?

B: Jane ate *some* of the cookies...

In the AM model for MAE, RFR is often described as using an L*+H accent with L-H% edge tones; an annotation first offered by Ward & Hirschberg (1985) and Hirschberg & Ward (1992). But L*+H is not the only rising accent[1] in MAE: both H* and L+H* are also rising accents. Hence, there exists not just a singular RFR but multiple RFR-shaped tunes.

Notably, there is some variation among researchers in which pitch accent(s) are used in a transcription of RFR; for instance, Büring (2003) describes a CONTRASTIVE TOPIC (CT) marking contour in AM terms as (L+)H*L-H%, which Constant (2012) claims should be treated as distinct from RFR which uses L*+H. Westera (2019) attempts to unify the meaning contribution of the

[1] Gussenhoven (2016) provides an alternative analysis of these accents (in the context of the nuclear tune) as falling accents, but this discussion is outside the scope of this work; see Barnes et al. (2021) for a discussion of this debate.

two tunes, where CT-marking is but one type of RFR's potential uses (see also Constant, 2014 ch. 5).[2] However, other researchers have argued that CT and RFR should be treated as distinct from one another (Wagner, 2012), with Göbel (2019, pp. 294–295) additionally observing that different types of CT-marking may be better understood as using different pitch accents (a distinction of L+H* versus L*+H). Briefly, in prior literature the potential distinctions between RFR-shaped tunes (characterized by differing pitch accents) are either intentionally unified, intentionally treated as distinct, or ignored for convenience. Given that there are competing and, at times, incompatible accounts of the **meaning** of RFR, the researcher-specific variation regarding the specific pitch accent, and its role in particular analyses, raises the question about whether distinctions in the intonational **form** are perhaps more relevant than what has been assumed.

There may also be reason to doubt a robust three-way categorical distinction between the RFR-shaped tunes. For MAE, it has been repeatedly shown that there is low inter-rater agreement when adjudicating between L+H* and H* (Pitrelli et al., 1994; Silverman et al., 1992; Syrdal & McGory, 2000), and the question of whether these accents comprise a single phonological category or two remains contentious (Ladd, 2008, 2022; Ladd & Schepman, 2003; Watson et al., 2008). In a study targeting the distinctions between the three rising pitch accents in MAE, Steffman et al. (2024) report substantial overlap in naïve speaker imitations of different RFR-shaped tunes that differ in pitch accent, and also that listeners were below chance at discriminating between L*+H from L+H* (c.f. prior imitation work from Pierrehumbert & Steele, 1989b and Dilley & Heffner (2013)).[3] An alternative view may thus see the three rising accents (and by extension, the three RFR-shaped tunes) as comprising a single phonological category that is subject to meaningful gradient variation in its phonetic expression (Ladd, 2008, pp. 154–156; Ladd, 2022, p. 252).

[2]Specifically, Westera (2019, p. 326) cites production data from Pierrehumbert & Steele (1989b) as evidence for a categorical distinction between a late-aligned versus an early-aligned bitonal pitch accent, but stops short of ascribing different meanings to each one. Westera notes that there may be differences, but that these are outside the scope of his account and are potentially paralinguistic (see also Gussenhoven, 2004 ch. 5).

[3]Similar results have also been found in German (which has a similar intonational phonology as MAE) in the context of CT-marking. Braun (2006) report inconsistent patterning of "thematic" pitch accents (L+H* and L*+H) in contrastive versus non-contrastive contexts in both production and perception with naïve participants. The researchers also found substantial disagreement among GToBI (the German implementation of ToBI, Baumann et al., 2000) annotators when annotating rising accents (see also inter-rater agreement reported in M. Grice et al., 1996).

This study examines whether listeners respond similarly to different RFR-shaped tunes, suggesting within-category phonetic variation in tune shape under a broad RFR class, or whether they elicit different responses, suggesting between-category variation attributable to the pitch accent specification. This research question is motivated further with a review of prior formal pragmatic accounts of RFR, which relate RFR to pragmatic alternatives. This connection has been investigated experimentally in the domain of SCALAR INFERENCE (SI) (Horn, 1972), examining the contrast between "neutral" (i.e., falling) intonation on the one hand and RFR on the other to adjudicate between accounts of RFR. The present work then uses SI as a testing ground to evaluate whether there are differences in offline interpretation and online processing using different falling and RFR-shaped tunes. The results obtained from a series of perception tasks (§4.2–§4.4) are discussed in relation to ongoing debates on SI and RFR as well as the implications for phonological structure. The goal of the present work is to examine the putative phonological distinctions and phonetic expressions of the RFR-shaped tunes to determine whether there exists a common core between them in terms of relating specifically to higher alternatives.

4.1.1 Rise-fall-rise

Within the AM framework, the meaning contribution of RFR has been intensely investigated over the past forty years (see Ward & Hirschberg, 1985 for a review of pre-AM descriptions as early as the 1930s). As such, there are a number of primary observations about the use of RFR in various discourse contexts that are uncontroversial and worth establishing at the outset.

First, RFR cannot be used out-of-the-blue with no context (Wolter, 2003), as in (3a) versus (3b). However, the contextual requirement here does not necessarily need to be a prior linguistic discourse as shown in (4).[4]

(3) Context: John has just entered his new office.

a. #John: The office is *warm*...

[4]# is used to denote infelicity. Note that (3a) would be felicitous with other intonational tunes such as a fall or a rise (see examples of H*H-H% described in Hirschberg & Ward, 1995). As before, *italics* denotes the nuclear-accented word and '...' denotes the phrase-final rise.

b. Maintenance Worker: We just fixed the AC, is the office okay?

John: The office is *warm*...

(4) Context: John has just arrived home to surprise his wife with take-out from her favorite restaurant, but walks in to find her visibly upset.

John: I got your *favorite*...

Second, RFR is generally more frequent in response to polar questions than *wh*-questions or declaratives (Ward & Hirschberg, 1985). In particular, RFR is frequently found when speakers "cannot, or do not wish to, commit themselves to direct responses" (Ward & Hirschberg, 1985, p. 769). Similarly, Wolter (2003) notes that RFR is frequently used when a speaker's answer is an incomplete answer to a question (see also the accounts of Wagner et al., 2013 and Westera, 2019).

RFR is also sensitive to scalar relations. Ward & Hirschberg (1985) note that for two scale values—an already discourse salient b_1 and a newly invoked value b_2—RFR is more frequent when b_2 is ranked lower than b_1 ($b_2 < b_1$, as in (5a)) than when $b_2 > b_1$ (as in 5b) or $b_2 = b_1$ (5c). While the use of RFR in (5b) and (5c) is pragmatically odd, RFR is perfectly acceptable if the scalar relationship targeted by RFR is modified via downward-entailing environments (as in 5d, see also Constant, 2012, p. 418 for an extended discussion) or by addressing an implicit question under discussion (QUD, Roberts, 1996) that is broader than what was originally asked (as explicated in (5e); see also Büring, 2003; Wagner et al., 2013; Westera, 2019).

(5) A: Is it cold outside?

a. B: It's *cool* outside... (cool<cold)

b. ?? B: It's *freezing* outside... (freezing>cold)

c. ?? B: It's *cold* outside... (cold=cold)

d. B: It's not *freezing* outside... (¬freezing≤cold)

e. B: It's *cold* outside... but it's not bad if that's what you're asking. (cold<bad)

In addition to the relative ranking of values, RFR is noted for being infelicitous when used with a scale endpoint as in (6) or with a value b_2 that is lower than b_1 along a measurement scale yet is

opposite in valence (Wolter, 2003, but see Göbel, 2019; Göbel & Wagner, 2023a for an extended discussion of valence asymmetry) as in (7).

(6) A: Did your friends like the movie?

#B: *All* of my friends liked it...

(7) A: Was the movie good?

#B: It was *bad*...

A number of formal pragmatic accounts have been proposed for RFR. Broadly, RFR has been described as conveying speaker UNCERTAINTY with respect to a scale (Hirschberg & Ward, 1992; Ward & Hirschberg, 1985); conveying a discourse strategy to relate the utterance to an alternative QUD (Büring, 2003; Wagner et al., 2013; Westera, 2019); or positing a close relationship between RFR and FOCUS (Krifka, 2008; Rooth, 1992) such that RFR imposes additional restrictions on the invoked set of focus alternatives (Constant, 2012; Göbel, 2019).[5] Despite the differences in accounts, a recurring theme between these accounts of RFR is a connection to alternatives: either focus alternatives (propositions), alternative questions (sets of propositions), or alternative speech acts. These alternatives are in some way secondary (Westera, 2019), broader (Wagner et al., 2013), or higher in terms of hierarchy (Büring, 2003) or scales (Göbel, 2019; Ward & Hirschberg, 1985).

As previously mentioned, while the focus of this literature is on "the" RFR tune, in actuality two RFR-shaped contours are described, differing in either an (L+)H* accent (Büring, 2003; Constant, 2012; Wagner et al., 2013) or an L*+H accent (Göbel, 2019; Hirschberg & Ward, 1992; Ward & Hirschberg, 1985; Westera, 2019). Some work takes a strong stance regarding the categorical distinction between the two tunes (Constant, 2012; Göbel, 2019; Hirschberg & Ward, 1992; Ward & Hirschberg, 1985) while others group the two together (Westera, 2019; Wolter, 2003) or are silent about the distinction between the two (Wagner et al., 2013).

[5]While not at issue in this work, RFR is also frequently discussed in relation to the marking of CONTRASTIVE TOPIC (CT). Some accounts equate RFR with CT-marking (Büring, 2003; Constant, 2014), but other work has suggested that the two phenomena are completely separate (Göbel, 2019; Wagner, 2012). Notably, Calhoun (2012) provides production evidence showing that neither H% nor the L-H% edge-tone configuration are robust markers of CT-marking (or "theme" in Calhoun's notation, vis-à-vis "rheme") for New Zealand English speakers. Other accounts offer a compromise wherein CT-marking is but one usage of RFR derived from a more general meaning contribution (Westera, 2019).

What stands out in the RFR literature is the disagreement about RFR in terms of not only intonational **function** but also intonational **form**. In parallel to the pragmatic literature is an additional debate regarding the robustness of the phonological contrasts between rising accents (Barnes et al., 2021; Dilley & Heffner, 2013; Iskarous et al., 2024; Ladd, 2022; Orrico et al., 2025; Pierrehumbert & Steele, 1989b; Steffman et al., 2024), which are implicated in distinguishing among three different RFR-shaped tunes. This disagreement on both sides motivates the question of whether claimed distinctions in the meaning function of different RFR-shaped tunes represent between-category variation or within-category variation with a common core in the meaning contributions of the three RFR-shaped tunes under a broad RFR class. Given the tacit assumption that RFR relates to alternatives, by some accounts higher alternatives specifically, a promising approach is to investigate RFR in the context of SI, which by definition invokes reasoning about higher alternatives.

4.1.1.1 Relating RFR to Scalar Inference

In the (Neo-)Gricean tradition, scalar inference is taken to arise via reasoning about what a speaker could have said but did not (H. P. Grice, 1975) with particular attention to pairs of lexical items that form a LEXICAL SCALE (i.a., Horn, 1972). Lexical scales are defined in terms of relative informativity, which is formalized using a relation of asymmetric entailment (Horn, 1972): for a scale $<X, Y>$, an utterance containing Y entails one containing X but not the other way around; hence, Y is the informationally stronger member of the pair and is often referred to as the *stronger* or *higher* scalemate while X is the *weaker* or *lower* scalemate. In 8, *some* is the weaker scalemate of the $<$*some*, *all*$>$ scale. On the definition of SI from Horn (1972), the listener reasons about the speaker's use of *some* in (8a) instead of the more informative scalemate *all*, as in (8b). Because the speaker did not utter (8b), the listener may infer that the speaker either does not know whether (8b) is true or that the speaker believes it to be false.[6] Thus, while the literal meaning of (8c) is

[6]Throughout this chapter, SI is taken to correspond to $\mathrm{K}\neg\phi$ (where K is an epistemic certainty operator and ϕ is a stronger alternative, following Sauerland, 2004), which is the meaning that is experimentally probed. This is the so-called secondary implicature described by Sauerland (2004). $\mathrm{K}\neg\phi$ is to be distinguished from the primary implicature ($\neg\mathrm{K}\phi$) and the ignorance inference ($\neg\mathrm{K}\phi \wedge \neg\mathrm{K}\neg\phi$)—these meanings are not directly probed in the experiments

compatible with (8b)—if Jane ate all the cookies, then via asymmetric entailment it is true that she ate some of the cookies—the listener is likely to arrive at the SI-enriched meaning in (8d), where the negation of the stronger alternative (8b) is incorporated into the enriched representation.[7] Such SI-enriched interpretations are possible for a variety of lexical scales, such as <*warm*, *hot*> or <*difficult*, *impossible*>.

(8) a. Jane ate some of the cookies. SENTENCE

b. Jane ate all of the cookies. ALTERNATIVE

c. Jane ate at least some of the cookies. LITERAL MEANING

d. Jane ate some but not all of the cookies SI-ENRICHED MEANING

Importantly, while the pragmatic SI-enriched interpretation is particularly robust for the <*some*, *all*> scale, it has been repeatedly shown that the likelihood of such SI enrichment (typically referred to as the "SI rate") varies across lexical scales—a phenomenon known as SCALAR DIVERSITY (Van Tiel et al., 2016). A large body of work in experimental pragmatics over the past two decades has sought to understand what factors contribute to the likelihood at which SI-enriched interpretations arise (i.a. Aparicio & Ronai, 2023; Doran et al., 2012; Gotzner et al., 2018; Ronai & Xiang, 2024; Sun et al., 2018). One important observation relevant to this work is that the likelihood of an SI-enriched interpretation is sensitive to prosodic factors such as whether the weaker scale term is uttered with a contrastive pitch accent (Thorward, 2009, see also Schwarz et al., 2007, Chevallier et al., 2008, and Zondervan, 2010 for disjunctive inferences with *or*). Importantly, recent work has shown that SI rates are also higher when the sentence containing a weaker scalar term is uttered with RFR (de Marneffe & Tonhauser, 2019; Ronai & Göbel, 2024). Accordingly, SI presents a testing ground for different accounts of RFR, with the benefit that the "higher alternative" involved in different accounts of RFR can be operationalized in terms of an

reported here. For experimental work relating RFR to the ignorance inference, see Buccola & Goodhue (2023).

[7]Terminologically, *scalar inference* refers **specifically** to the inference regarding the negation of the higher alternative—not any inference that makes use of a lexical scale. So, *Jane ate some but not all of the cookies* is an instance of SI, but *Jane ate all, not merely some, of the cookies* is not an instance of SI. Relatedly, this work will continue to use the term *scalar **inference*** rather than *scalar **implicature***, as the focus is primarily about the listener's interpretation and not the speaker's intended meaning.

alternative containing the higher scalar term.

The different pragmatic accounts of RFR described in the previous section allow different predictions for SI. Accounts vary in whether they predict a lower likelihood of SI (Büring, 2003; Ward & Hirschberg, 1985),[8] a greater likelihood of SI (Constant, 2012; Göbel, 2019), or do not make clear predictions one way or the other (Wagner et al., 2013; Westera, 2019). Yet, on the definition of SI from Horn (1972), the listener can arrive at the SI-enriched interpretation based solely on the propositional content of the utterance: reasoning about the speaker's use of *some* in *Jane ate some of the cookies* instead of the more informative scalar alternative *Jane ate all of the cookies* will lead them to the SI-enriched interpretation *Jane ate some **but not all** of the cookies.* Thus, intonation is not a **necessary** component for SI computation. However, there are experimental results showing that RFR makes this enrichment more likely.

4.1.2 Experimental Work on RFR and Scalar Inference

Due to the connection between RFR and higher alternatives, recent empirical work on RFR has focused primarily on how RFR relates to SI. For instance, in a perception study on RFR, de Marneffe & Tonhauser (2019) investigated how RFR contributes to the interpretation of indirect answers to polar questions like (9).

(9) Mike: Was your hike exhausting?

Julie: It was *strenuous*....

In their study, participants were tasked with responding whether Julie means that her hike was exhausting using a 7-point scale from *Definitely No* to *Definitely Yes*. Here, if the participant arrives at the SI-enriched interpretation *The hike was strenuous but not exhausting*, then they should be more likely to give lower ratings (i.e., No, Julie did not mean her hike was exhausting). They found that utterances made with RFR were more likely to receive lower ratings (=greater likelihood of

[8] While the uncertainty account from Ward & Hirschberg (1985) posits three different types of uncertainty, this work follows de Marneffe & Tonhauser (2019) and Ronai & Göbel (2024) in focusing primarily on Type III: uncertainty about the choice of some value on the scale. Here, uncertainty about, e.g., whether *cool* is close enough to *cold* (for the purposes of the conversation) predicts a lower likelihood of SI by virtue of requiring the truth of the higher *cold* to remain unresolved; this is incompatible with SI, which renders the higher alternative false.

SI) compared to utterances made with falling intonation. They take their results as evidence that prosody can serve as a cue to the speaker's intended interpretation of their utterance, where RFR "strengthens the degree of belief in the scalar implicature over the neutral contour" [12], which is not predicted by Wagner et al. (2013) or Ward & Hirschberg (1985). However, because participants' ratings were generally high (with the most likely response for both intonation conditions being 5/7, the option labeled *Perhaps Yes*, de Marneffe & Tonhauser (2019) take these results to suggest that accounts where the higher alternative is strictly negated (one possibility of Constant, 2012) are too strong.

Buccola & Goodhue (2023) also look at SI with regard to RFR and falling intonation, but additionally pit an SI-enriched interpretation up against the mutually exclusive uncertainty interpretation (termed an IGNORANCE INFERENCE, or II).[9] Their work is situated within a theory that the grammatical marking of uncertainty is obligatory (Buccola & Haida, 2019) and so if RFR serves as an intonational marker of uncertainty, its usage should be favored over alternative tunes in situations where the speaker needs to convey uncertainty. In their perception study, participants are asked to listen to two utterances containing the quantifier *some*, e.g., *Some of them ate dinner*, where one rendition uses falling intonation and the other uses RFR. Participants made a binary choice to judge which tune was associated with the SI-enriched interpretation and which tune was associated with the II-enriched interpretation.[10] Using this kind of intonation-interpretation mapping paradigm, they found that RFR was more likely to be mapped to the II-enriched interpretation than the SI-enriched interpretation, although there was a general preference to compute SI regardless of intonation. However, it must be noted that while the use of RFR may be compatible with either an SI- or II-enriched interpretation, the use of falling intonation in their study was only compatible with the SI-enriched interpretation. Accordingly, the preference for mapping RFR to the

[9]Note that Ward & Hirschberg (1985) use "uncertainty" to express a 'lack of speaker commitment,' which encapsulates both $\neg K\phi$ and $K\neg\phi$ in felicitous examples of RFR that do not make use of entailment-based scales. However, later work on RFR from other researchers, at least in the context of SI with entailment-based scales, view "uncertainty" as not covering the $K\neg\phi$ meaning.

[10]An example of response choices A and B in their study is: (A) is *In version 1, Mason thinks that not all of the guests ate dinner, and in version 2, Mason isn't sure whether or not all of the guests ate dinner.* and (B) *In version 1, Mason isn't sure whether or not all of the guests ate dinner, and in version 2, Mason thinks that not all of the guests ate dinner.*

II-enriched interpretation may have arisen due to falling intonation necessarily needing to map to the SI-enriched interpretation.

Lastly, Ronai & Göbel (2024) report results from a dual production and perception task using question-answer dialogues like (10). In their STRONG condition, the question used the stronger lexical scalemate while in their SAME condition, the weaker scalemate is used in the question. For the production portion, participants are asked to record their response to Emma's question using the given sentence. Following their production, participants are asked an SI-probing question, where a response of *Yes* indicates that they computed the SI-enriched interpretation and a response of *No* indicates that they did not.

(10) a. Emma: Was the winner happy? SAME

b. Emma: Was the winner ecstatic? STRONG

You: She was happy.

Given your response, do you think Emma would conclude that the winner is not ecstatic?

Their prediction was that RFR should be more likely to be used in the STRONG condition compared than the SAME condition, where it would generally be considered infelicitous. Their production results show that RFR was typically used in the STRONG condition as opposed to the SAME condition; however, falling intonation was still the predominant intonation used by participants. The Yes/No responses to the SI-probing question were then conditioned on the participant's choice of intonation on each trial. They found that SI rates were higher when participants used RFR than when they used falling intonation in the preceding production portion of the task. In a followup inference task with no production portion, they again found that SI rates were higher when RFR was used compared to when falling intonation was used.

The authors take these results as evidence in favor of the salience account proposed by Göbel (2019) and Göbel & Wagner (2023a). In this account, RFR makes the higher alternative more salient by adding a presupposition that such a higher alternative exists, in turn making SI-enrichment more likely.[11] The result that such SI-enrichment is more likely is taken to be incom-

[11]On some views of SI, the inference is computed automatically if the alternative is "sufficiently salient" (Bott &

patible with accounts that require the truth of the higher alternative to remain unresolved (Ward & Hirschberg, 1985 and Wagner et al., 2013). They note that the results are nonetheless compatible with the accounts of Constant (2012) and Westera (2019), as these could predict either an increase or a decrease in SI computation.

Outside of narrowly investigating SI, empirical work has shown that RFR with higher pitch range is more likely to be interpreted as conveying that the speaker is incredulous rather than uncertain (Hirschberg & Ward, 1992) and that RFR, compared to falling intonation, is more likely to be rated as expressing speaker uncertainty and that the speaker is trying to "insinuate something above and beyond what is literally asserted" (Wagner et al., 2013, p. 148). Moreover, Wagner et al. (2013) find that when given complete-answer (11) or partial-answer (12) contexts, participants were more likely to use RFR in the partial context rather than the complete context.[12]

(11) Complete-Answer Context: (Wagner et al.'s 15)

Q: Is Bill coming to the party?

?? A: *Bill* is coming...

(12) Partial-Answer Context: (Wagner et al.'s 16)

Q: Is either Bill or Susan coming to the party? A: *Bill* is coming...

In a rating study focusing on valence asymmetries, Göbel & Wagner (2023a) finds that RFR is more felicitous with negative statements in response to positive statements than the reverse, but that this asymmetry is not present in responses to questions. Because the focus of the present work is primarily on the use of RFR in response to questions, questions regarding valence asymmetry as it relates to RFR will be set aside.

Chemla, 2016; Bott & Frisson, 2022; Rees & Bott, 2018 but c.f. Marty et al., 2024). *Salience*, however, is somewhat loosely defined in terms of cognitive activation and shows different degrees of success in increasing SI rates across studies where salience is modulated in different ways (De Carvalho et al., 2016; Schwarz et al., 2016).

[12]Wagner et al. (2013) does not specify which word in these examples bear the nuclear pitch accent, but it is assumed here (based on their provided judgments) that it should be *Bill*.

4.1.2.1 Limitations of Previous Work

Empirically, RFR has been investigated through either rating studies (de Marneffe & Tonhauser, 2019; Göbel & Wagner, 2023a; Wagner et al., 2013) or forced choice paradigms (Buccola & Goodhue, 2023; Ronai & Göbel, 2024). The rating tasks show a potential response bias across experiments (described below) while the more recent forced choice paradigms show either mixed results in the matching paradigms (Buccola & Goodhue, 2023) or rather straightforward results from inference judgment paradigms (Ronai & Göbel, 2024).

Focusing on the rating studies, there is an apparent recurring response bias where either the majority of responses or the aggregate measure of central tendency lies at just above the midpoint of the scale. In de Marneffe & Tonhauser (2019), the results show that for both falls and RFR, the vast majority of responses are 5/7—the point just above the middle of the scale (=4).[13] In the four rating tasks presented in Wagner et al. (2013), while it is difficult to tell what the exact distribution of the results is from the boxplots provided in the paper, the medians of all intonation conditions were generally located slightly above the midpoint.[14] In Göbel (2019), results from naturalness ratings (from 1-6) are presented in bar graphs which do not show the underlying distributions, but the mean ratings for both falls and RFR in their matching condition range between 4 and 5 out of 6, where the scale midpoint would be 3.5, with RFR being credibly lower (though the effect size is not reported). Taken together, these results suggest independent evidence for a response bias where participants tend to select the option just above the midpoint that corresponds to, roughly, the *Perhaps Yes* or *Somewhat Confident/Relevant/etc.* response option regardless of intonation.

In the context of these rating tasks, it is difficult to determine how substantial the differences between intonation conditions are when participants overwhelmingly use response options corre-

[13] In unpublished work (https://github.com/thegricean/speaker_prosody) from Degen, Tomlinson, and Waldon that builds on de Marneffe & Tonhauser (2019), a 0-100 sliding scale was used instead of a 7-point scale. The results showed the same pattern as de Marneffe & Tonhauser (2019), where the mean ratings for RFR and "neutral" prosody hovered at or just below 75% (note that 5/7=71.4%).

[14] These rating tasks queried Acceptability, Speaker Confidence, Relevance, and Insinuation. The ratings are supposed to be from 1-7, but some of the results are shown on a scale of 0-8, making it difficult to tell whether a rating of 6 should be interpreted as a 6/8, a 6/9, or a 5/7. For their Insinuation question, unlike the other three questions, the ratings are generally expected to be on the low end of the scale rather than the high end. Accordingly here, the medians appear to be slightly below the midpoint.

sponding to *Perhaps Yes*. There is a broader point to be made here about these kinds of offline interpretation tasks, related to ACCOMMODATION. Intonational meaning (in MAE at least) is often difficult to assess experimentally because participants can accommodate multiple interpretations for multiple tunes and vice-versa—i.e., a (oftentimes probabilistic) many-to-many mapping between intonational form and function (Cole, 2015; Hirschberg, 2017; Roettger et al., 2019; Sostarics & Cole, 2023a, see also chapter 2 of this thesis though c.f. arguments by Arvaniti et al., 2024 in the context Greek, a non-Germanic language). This discussion is not to suggest that prior results are not meaningful, but rather that the magnitude of the perceived contrast between intonational tunes may be masked due to accommodation.[15] Given that such a response bias complicates identifying between-category differences among broad tune classes, like RFR or Fall, rating tasks do not seem promising for identifying within-category differences among these broad tune classes. The inference task, at least as applied to SI calculation, seems more fruitful as a measure of offline interpretation.

4.1.3 Goals

Given the connection between RFR and higher alternatives, one might ask whether this relationship is restricted to a single RFR or whether all RFR-shaped tunes share this connection as a common core (see also Constant, 2014, p. 279 who explicitly claims the three share a common CT use). Though all accounts rely on some notion of ALTERNATIVE, some accounts (Constant, 2012; Göbel, 2019) concern themselves specifically with the set of focus alternatives. Given this, we might expect that these connections would only hold for the bitonal accents (Pierrehumbert & Hirschberg, 1990)—or even more specifically, the L*+H accent alone (Göbel, 2019). Going further, if these effects are generally related to the choice of bitonal pitch accent, would we expect similar effects for falls that share these pitch accents? Or is it necessary to instead consider the holistic tune?

[15] One may object that these results may be interpreted instead as a binary choice with varying degrees of confidence, where *Perhaps Yes* reflects lower confidence than *Definitely Yes*. However, in response to the question *Does Julie mean that her hike was not exhausting?*, it would be contradictory to respond *Definitely yes, but she might not* whereas it is not contradictory to say *Perhaps yes, but she might not.* That is, a *Perhaps Yes* response is not all that dissimilar to a *Maybe* response.

The goal in this work is to unpack the typical labels of "RFR" and "neutral" (i.e., falling) intonation and determine whether RFR-shaped tunes that differ in the specification of the pitch accent behave **similarly** or **differently** from one another in eliciting listener responses. SI is used as a testing ground to operationalize what is meant by HIGHER ALTERNATIVE—here, it is the informationally stronger member of the lexical scale (Horn, 1972) vis-à-vis what will be referred to as the LOWER ALTERNATIVE (traditionally referred to as the weaker member of the scale). For example, in <*cool*, *cold*>, *cold* is the higher alternative and *cool* is the lower alternative. In addition to examining SI using an offline judgment task, this work also asks whether differences in RFR-shaped tunes are apparent in online processing using cross-modal priming with lexical decision, which has previously been used to probe the timecourse of processing for focus alternatives (see Gotzner & Spalek, 2019 for a review). An overview of the relevant background and findings for this work is given in the next section, which will help to provide some common ground regarding (1) constraints for the experimental materials used in this study (discussed in Section 4.2.1) and (2) terminology for discussing the predictions. Additionally, by capitalizing on RFR's purported connection to higher alternatives and by relating **scalar** alternative processing to **focus** alternative processing, there is an opportunity to address whether there is a distinction between the two; Section 4.1.4.1 will thus motivate a complementary research question related to this.

4.1.4 Cross-Modal Priming, Alternative Activation, and Intonation

The cross-modal lexical decision paradigm (i.a., Braun & Tagliapietra, 2010; Husband & Ferreira, 2016) is used in this work to probe the activation status of higher and lower alternatives. In this paradigm, a sentence containing a PRIME word, such as *The museum thrilled the sculptor when they called about his work* is presented auditorily (*sculptor* is the PRIME). After some delay a TARGET, either a word (*painter*) or a non-word (*fronk*), is then presented in text on a screen.[16] Participants are asked to judge whether the TARGET is a word or not a word of English. The hypothesis for

[16]The delay between the offset of the auditory PRIME and the onset of the visual TARGET is referred to in this literature as the STIMULUS ONSET ASYNCHRONY, or SOA. For example, a 0ms SOA indicates that the visual target appears immediately after the offset of the auditory prime, while a 750ms SOA indicates a 750ms delay.

this paradigm is that lexical retrieval is easier/faster for lexical items that have higher levels of cognitive activation compared to lexical items that have lower levels of activation at the time of retrieval. Hence, participants' REACTION TIME (RT) when correctly identifying a real-word target as a word can index the activation status of particular lexical items (Braun & Tagliapietra, 2010; Husband & Ferreira, 2016; Swinney et al., 1979; Tabossi, 1996). This paradigm is often used to evaluate how the relationship between the PRIME and the TARGET is modulated in the presence or absence of different linguistic features such as intonation (see Gotzner & Spalek, 2019 for an extensive review).

In a study investigating the timecourse of contrastive focus processing as cued by intonation, Husband & Ferreira (2016) found that the intonation with which the sentence containing the auditory PRIME (e.g., *sculptor*) is uttered has different effects depending on whether the TARGET can serve as a contrastive focus alternative (*painter*) beyond being merely semantically associated (*statue*). The results of that study showed that both SEMANTIC ASSOCIATES (the CONTRASTIVE *painter* and NON-CONTRASTIVE *statue*) were initially facilitated relative to semantically UNRELATED targets (*register*) regardless of whether "neutral" H* or "focus-marking" L+H* was used. However, at a later timepoint, and only in the focus prosody condition, the CONTRASTIVE associate *painter* (i.e., the focus alternative) alone continued to show facilitation while the NON-CONTRASTIVE associate *statue* was deactivated, suggesting a process of active suppression of non-contrastive semantic associates. The results, schematized in terms of activation level, are shown in Figure 4.1.

4.1.4.1 Complementary Aims Regarding the Processing of Scalar Alternatives

The finding from Husband & Ferreira (2016) that contrastive focus alternatives behave differently from mere semantic associates is situated within a larger enterprise regarding the processing of alternatives. An important question in this literature is whether alternatives are represented and accessed in processing. For instance, if the role of a focus-sensitive particle like *only* is to exclude alternatives to the constituent it associates with, then those alternatives must be activated to

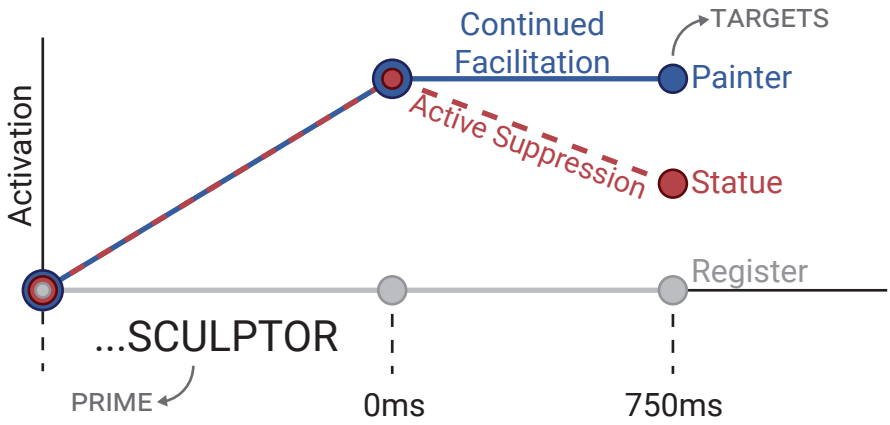


Figure 4.1: Schematic of activation for contrastive associate (*painter*), non-contrastive associate (*statue*), and unrelated word (*register*) at two timepoints. Participant RT is inversely related to the activation of the lexical item (more activated=faster/lower RT).

be excluded (Gotzner, 2019; Gotzner & Spalek, 2017; Lacina et al., 2024). In a review of empirical work on alternative activation, Gotzner & Spalek (2019, p. 12) conclude that "the general effect of focus [..] is **facilitatory**: If an effect of focus on the availability of focus alternatives is present at all, it is **priming**, increased fixations or improved recognition performance. Therefore, focus marking **increases the availability of alternatives** to the focused element" (emphasis mine). Thus, there is ample work on the psychological reality of **focus** alternatives and their processing correlates, but does this work extend directly to **scalar** alternatives?

Whether there is in fact a distinction to be drawn between the alternatives involved in SI and those involved in focus is an open question. For example, in the sentence *Jane ate some of the cookies*, one alternative to be excluded for SI would be *Jane ate all of the cookies*; yet, if *some* were placed in narrow focus, or if the sentence used the focus particle *only* as in *Jane ate only some of the cookies*, we would again exclude *Jane ate all of the cookies* via an exhaustivity operator (Chierchia, 2004). Indeed, in contrast to the view that the stronger alternative for SI is lexically specified, Fox & Katzir (2011) argue that the alternative-generating process for alternatives related to focus and SI is the same. Such a theoretical proposal might indicate a common core in the processing of alternative exclusion in both SI and focus (see also Gotzner & Romoli, 2022 for a related overview) that does not require the notion of a lexical scale to account for SI. More broadly, Post-Gricean

(PG) theories of pragmatic meaning (e.g., Relevance Theory, Wilson & Sperber, 2006) similarly capture SI derivation without reference to lexical scales—hence dispensing with scales entirely—in contrast to Neo-Gricean (NG) accounts of SI (Horn, 1972), which ascribe special importance to reasoning about lexical scales. Taken together, we can observe that there are competing accounts about SI derivation as well as whether scalar alternatives and focus alternatives should be treated as distinct from one another.

As mentioned, there is ample work on the processing of focus alternatives; yet analogous evidence for the processing of scalar alternatives (if there is to be a distinction) is more limited and mixed but suggests a similar facilitatory effect. De Carvalho et al. (2016) report an asymmetry in priming using a subliminal priming task (in French with words in isolation) such that lower alternatives like *some* prime higher alternatives (like *all*) more than the reverse (c.f. Schwarz et al., 2016, who report no effect of subliminal priming on SI likelihood). They take their results to reflect the psychological reality of lexical scales, as otherwise no asymmetry in priming would arise. In a similar task, Ronai & Xiang (2023) only found priming of the higher alternative when the lower alternative was used in a sentence but not when presented in isolation. Moreover, this priming persists even when *only* (an overt focus-sensitive operator) is used. The authors take this as evidence that priming is most likely driven by the inferential process involved in relating the weaker alternative to the stronger alternative (i.e., requiring sentential interpretation).[17] Building on these two studies, Lacina et al. (2024) find evidence that a stronger alternative of a scale (*filthy*) is not activated when its weaker scalemate (*dirty*) is sententially presented under negation (e.g., *Zack's carpet was not dirty*), which they take as evidence that *filthy* is not primed when it is not reasoned

[17] De Carvalho et al. (2016) were largely focused on adjudicating between NG and PG accounts using priming and took their results to be support for a (rather strong) NG account where the interpretation of weaker scalar items requires accessing the stronger alternatives but not vice versa—an asymmetry which they argue is incompatible with PG accounts. The authors used subliminal priming under the implicit assumption that such an asymmetry would be apparent even when the scalar item is not in a sentence, e.g., rapid visual presentation of merely the word *some* would necessarily activate *all* even outside of interpreting a sentence. However, Ronai & Xiang (2023) warn that there are ways in which priming of the stronger alternative could in fact be accounted for without recourse to accessing the stronger alternative specifically via spreading activation through shared semantic features—hence the asymmetry may be epiphenomenal and not strictly incompatible with PG accounts. The adjudication between NG and PG accounts is largely outside the scope of this work, but highlights the at-issue question of whether scalar alternatives are psychologically real in the same way focus alternatives have been shown to be. For further discussion, see Ronai & Xiang (2023) and Lacina et al. (2024).

about (because it is no longer informationally stronger). While Ronai & Xiang (2023) and Lacina et al. (2024) attribute their results largely to the presence of implicature processing, they do not additionally test whether such priming between members of a lexical scale is asymmetric—i.e., they don't test whether a stronger alternative like *filthy* primes *dirty*, where the former asymmetrically entails the latter.

One limitation of previous priming studies investigating whether scalar alternatives are the same as focus alternatives is that they do not actually probe whether or not SI was calculated during the lexical decision task. For instance, while the asymmetry found by De Carvalho et al. (2016) is in line with their view on SI (i.e., that SI arises automatically from lexical items and thus drives the asymmetry by definition), we know from work on scalar diversity that SI is a pragmatic inference that does not arise 100% of the time and that the probability of such an inference is not uniform across lexical scales (Doran et al., 2012; Van Tiel et al., 2016). As a result, it is unclear whether faster RT on any particular trial is strictly the consequence of SI computation or not. In recent work trying to fill this gap, Lacina & Gotzner (2024) report a between-experiment correlation between SI rates from the inference task results from Gotzner et al. (2018) and a text-based priming task where the activation status of a higher alternative such as *hot* was probed following a sentence like *It is warm* (where *warm* is the corresponding lower alternative of the scale, c.f. their UNRELATED condition *It is lucky*). The between-experiment results showed that scales with higher SI rates were in fact associated with **slower** RT in the priming task—an unexpected result that runs counter to prior assumptions in other priming tasks (De Carvalho et al., 2016; Ronai & Xiang, 2023). Thus, it is worth addressing whether the relationship between the activation of scalar alterantives (tested via priming) and the calculation of SI can be more explicitly probed within the same task.

The present work will report results from an inference task probing SI calculation (Exp. 1) and multiple lexical decision tasks probing the activation status of scalar alternatives involved in SI calculation (Exp. 2-3). Additionally, a novel dual task paradigm (Exp. 4) is used to address whether priming results in lexical decision are present or absent depending on whether SI is actually calculated.

As previously described, the generalization for RFR based on prior work is (broadly speaking) that it is related in some way with higher alternatives. If we expect scalar alternative activation to behave similarly to focus alternative activation, then we would predict that the use of RFR with a weaker alternative should facilitate the stronger alternative beyond what would normally be expected by mere semantic priming. For example, with a lexical scale like <*cool*, *cold*>, *cool* and *cold* should prime one another due to mere semantic association, but we would predict that *cool* uttered with RFR additionally facilitates *cold* while *cold* uttered with RFR **does not** additionally facilitate *cool* due to the connection between RFR and higher (not lower) alternatives. Testing this prediction requires probing the activation status of the higher alternative *cold* and the lower alternative *cool*—the exact piece missing from Ronai & Xiang (2023) and Lacina et al. (2024). Because of this directionality inherent in the analysis of RFR, which drives the experimental design of this work, there is a unique opportunity to situate this work in relation to ongoing discussions on the potential distinction between scalar alternatives and focus alternatives.

To motivate the complementary research question of whether scalar alternatives differ from focus alternatives further, consider again the scale <*cool*, *cold*>. Both scalemates can be contrastive focus alternatives to one another. However, these are **also** related to one another via asymmetric entailment (Horn, 1972); that is, they are also scalar alternatives. Might we expect the scalemate relation to afford a distinction in processing above and beyond the contrastive alternative relation? This question is complementary to the main goal of this work (assessing the effect of **intonation** on the processing of higher versus lower alternatives) and focuses solely on the **lexical** relationship between the scalemates. That is, *cool* and *cold* are scalemates **regardless** of the intonation with which *cool* is uttered in much the same way that *cool* is semantically related to *cold* regardless of the intonation with which *cool* is uttered.

When considering the possible targets in a cross-modal lexical decision task following a prime like *cold*, we can consider the scalar alternatives as a further distinction to the contrastive condition in Husband & Ferreira (2016). Taking a hierarchical view, all scalar alternatives can be focus alternatives, but not all focus alternatives are scalar alternatives. The hierarchy of the possible tar-

get set, including scalar alternatives in addition to the conditions presented in Husband & Ferreira (2016), is summarized in Figure 4.2.[18]

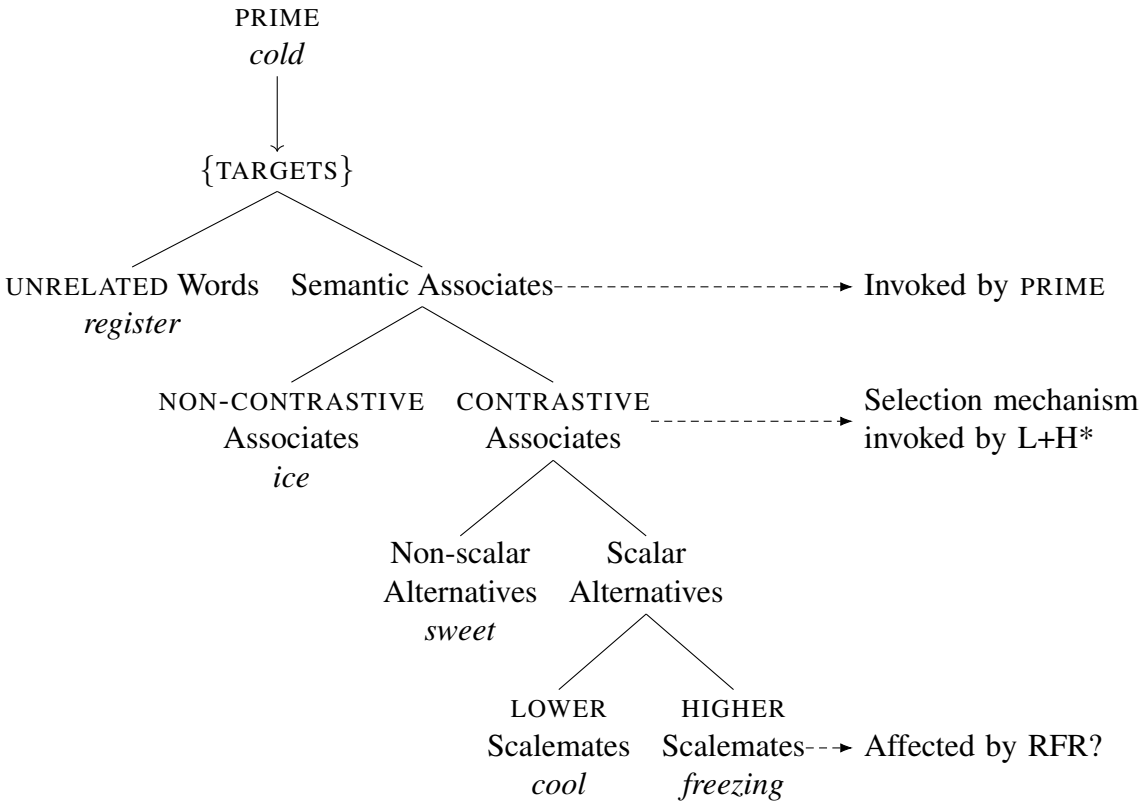


Figure 4.2: Hierarchical breakdown of the potential set of target words in relation to a given prime word. Examples of words for each subset are given in *italics*.

Before proceeding to the experimental materials for this work, the research questions can be summarized follows.

1. Do RFR-shaped tunes that differ in the specification of the pitch accent behave similarly or differently from one another in offline interpretation and/or online processing in the context of SI?
2. Do scalar alternatives behave differently from focus alternatives in processing, abstracting away from intonational differences?
3. Is there an asymmetry in the processing correlate of RFR such that it **specifically** targets the higher alternative (e.g., to negate it as a result of SI), and not the lower alternative?

[18]Missing from this hierarchy is the class of antonyms to the prime word, such as *hot*. This can be seen as a distinction between opposite/same valence nested under the scalar alternatives node or as a further distinction within the lower scalemates, allowing for an ordering relation to operate at the full measurement scale. Ultimately though, antonyms are outside the scope of this work (see also notes by Wolter, 2003), but see Lacina et al. (2024) for a text-based priming study with antonyms.

4.2 Norming Task for Written Materials

Prior to conducting the main inference and lexical decision tasks, a norming task was conducted using only written materials, which are described below. The norming task serves three purposes. First, it helps to ensure that the discourses used in the inference and lexical decision tasks are not deemed unnatural by participants, which would preclude questions about pragmatic interpretation. Second, it helps to determine which items should not be included in the subsequent experiments, which avoids having to record them. Lastly, it serves as a sanity check for whether the SI-enriched interpretation is available.

4.2.1 Written Materials

This work uses question-answer dialogues such as (13). These dialogues use indirect answers similar to de Marneffe & Tonhauser (2019) but, unlike that study, the question does not explicitly mention the higher alternative.[19] For a polar question like (13), the most straightforward answer would be either Yes (I did) or No (I did not).

(13) Mary: Did you come up with anything for the last problem on the exam?

a. John: That one was difficult (lower scalemate)

b. John: That one was impossible (higher scalemate)

The indirect answers shown in (13a-b) may convey either an implicit Yes or an implicit No response depending on how the participant reasons about the answer within the context. For example, given (13a), the listener may derive the SI that the problem was difficult but not impossible, and thus conclude that John was able to come up with an answer. Alternatively, the listener may not derive the SI-enriched interpretation and conclude that the problem was too difficult for John to answer. The key point here is that while a relevance implicature is needed for the answer to be judged felicitous in context, SI calculation is not necessary. If SI were necessary to establish

[19] In anticipation of the upcoming cross-modal lexical decision task, if the goal is to probe the activation status of *cold* given an auditory prime of *cool*, then *cold* cannot also be explicitly included in the question context itself, as this would likely oversaturate any potential effects of intonation.

relevance, then this would artificially inflate SI rates and likely mask any potential effect of intonation. In contrast to examples like (14), where it is intuitively far more difficult[20] to establish the relevance of John's answer to Mary's question, the critical items should be judged as contextually more acceptable. If the critical items in this study are deemed similar to irrelevant answers as in (14), then this would serve as a signal to rewrite or omit the item from future experiments.

(14) Mary: Did you do the extra readings for class?
#John: There used to be a Burger King

A second coder uninvolved with the project and I coded each critical question-answer pair by whether a literal interpretation is available, e.g., *That one was difficult and even impossible*, and whether the SI-enriched interpretation is available, e.g., *That one was difficult but not impossible*. Inter-rater agreement for the availability of each interpretation was assessed via Gwet's AC1, showing very strong agreement (literal interpretation AC1=.98, SI-enriched interpretation AC1=1). Both coders also coded whether, given either the literal or SI-enriched interpretation, the conveyed answer with that intended interpretation was either a Yes or a No (rated on a 5 point scale). The correlation between implicit answer ratings was assessed using Kendall's τ_b, showing strong correlations between ratings (literal interpretation $\tau_b = 0.70$, SI-enriched interpretation $\tau_b = .66$). Where there were disagreements, the trials were rewritten until they passed both criteria with both coders. Only a small number of trials (n=3) did not pass both criteria with both coders. All of the scales were nonetheless included in the norming task.

Following the same structure as the critical items, 60 filler items were written.[21] The norming task uses 30 of these filler item question-answer pairs where the answer was irrelevant by shuffling the questions and answers as in (14).[22]

[20]"Difficulty" will be operationalized in terms of surprisal in the next subsection.

[21]Nine were adapted from materials in Domaneschi et al., 2017, 13 were adapted from a public repository of questions used on the dating site OKCupid, and the rest were written from scratch.

[22]To ensure that the resulting dialogues were unnatural, the overall surprisal (i.e., the inverse of the probability of the answer given the question) of each possible question-answer pair using GPT-2 Large was calculated. Then, the pairing of questions to answers is treated as a stable marriage problem, which is solved using the Gale-Shapley algorithm implemented in the matchingR R package (Tilly & Janetos, 2021) to pair each question with an answer that maximizes the total overall surprisal of the 30 pairings.

All items were written under the consideration that they would be eventually recorded. Thus, all of John's answers were written to be a single prosodic phrase approximately 6 syllables long when the lower scalemate was used. The sentence lengths for the critical and filler items[23] ranged from 4 to 9 syllables and the averages for each item set were not significantly different from one another based on a Conway-Maxwell-Poisson model (Sellers et al., 2023) for underdispersed count data[24] ($\hat{\beta} = 0.10, SE = 0.19, z = 0.52, p = 0.606$).

4.2.2 Procedure

Participants were told that they would be reading short dialogues between Mary and John, where Mary would ask a question and John would respond with an indirect answer.[25] The task is to first rate how acceptable John's response is as an answer to Mary's question using a 6-point likert scale from 1=Completely Unacceptable to 6=Completely Acceptable; in these examples, the prediction is that (13) will receive high ratings and (14) will receive low ratings. After rating the dialogue, participants are asked a Yes/No comprehension question (described next). Participants were instructed to give their spontaneous responses and that they should not spend too much time thinking about any one trial.

Because the items need to be normed both when the lower scalemate (e.g., *difficult* as in 13a) is used **and** when the higher scalemate is used (e.g., *impossible* as in 13b), the comprehension question differs slightly depending on which scalemate is shown in the dialogue. Examples of both conditions are shown in (15) and an example of a filler item is shown in (16).[26] If participants are shown a dialogue that uses the lower scalemate, they are asked whether they would conclude

[23]To briefly foreshadow the future experiments, a third item set that adapts the items from Husband & Ferreira (2016) was also created. These items will be discussed in more details in Section 4.4.1, but relevant here is that they had the same polar question/indirect answer structure as the critical and filler items. They were also held to the same standard for the length of the prosodic phrase and were not significantly different from the critical trials ($\hat{\beta} = 0.16, SE = 0.19, z = 0.86, p = 0.390$)

[24]This model accounts for underdispersion but the same result holds regardless of choice of linear regression, poisson regression, or negative binomial regression.

[25]Note that Mary always asked the question and John always responded, gender presentation is not manipulated in this experiment nor in subsequent experiments.

[26]The differences between conditions are emphasized in boldface but are not bolded in the actual task. Additionally, punctuation was not included with John's responses so as to not influence the implicit prosody participants ascribe to John's responses.

...*not impossible* (as in 15a), modeled after inference tasks used in Van Tiel et al. (2016) and Ronai & Göbel (2024)). If participants derive the SI-enriched *difficult but not impossible* interpretation, participants would respond Yes, otherwise they would respond No. Conversely, if the participant is shown the higher scalemate, as in (15b), they are asked whether they would conclude ...*not merely difficult*. For expository ease, the rate at which participants say Yes to these "merely" questions will be referred to as the MERELY INFERENCE rate (MI rate, analogous to SI rate).[27]

(15) Mary: Did you come up with anything for the last problem on the exam?

a. John: That one was **difficult**

Prompt: Would you conclude from John's response that the problem was **not impossible?**

b. John: That one was **impossible**

Prompt: Would you conclude from John's response that the problem was **not merely difficult?**

(16) Mary: Did you do the extra readings for class?

John: There used to be a burger king

Prompt: Would you conclude from John's response that he did not do the readings?

The norming task included 102 total trials: 36 critical trials used the lower scalemate, 36 critical trials used the higher scalemate, and 30 trials were fillers. Critical item conditions were counterbalanced between participants.

4.2.3 Results

Undergraduate students at Northwestern University participated for course credit (n=66). Participants were excluded if they self-reported as not being naive MAE speakers (n=15) or self-reported

[27]The MI rates are not at issue in this work. Rather, the inclusion of the "merely" questions is driven solely by methodological considerations so that both higher and lower conditions have similar trial structures involving a comprehension question **with negation** of the relevant scalemate. Accordingly, the MI-rate results are entirely exploratory.

hearing (n=1) or vision (n=2) deficits. A total of 48 participants were included for further analysis (24F, 20M, 4 Other; mean age 19.9).

The empirical results for the acceptability ratings are shown in Figure 4.3. From the figure, it is clear that the fillers were, as expected, deemed largely unacceptable while the critical items were judged as acceptable. A by-item breakdown of the ratings is shown in Appendix C.3, Figure C10. There were a few critical items that received a relatively high ($> 15\%$) proportion of low ratings (rating=1-3), which were removed from subsequent experiments.

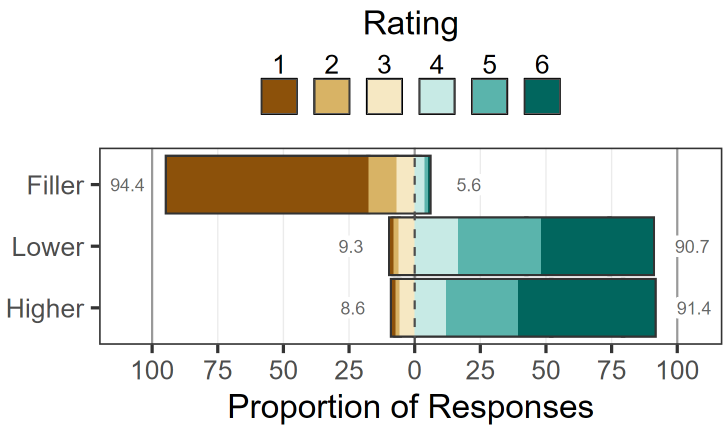


Figure 4.3: Norming task rating results. Proportion of ratings (1=Completely Unacceptable, 6=Completely Acceptable) for each condition. Lower/Higher indicates whether lower/higher alternative was used in John's answer. Hence, the Lower condition probes SI while the Higher condition probes MI. Numbers give the proportion of responses falling on the low (1-3) versus high (4-6) sides of the scale. Bars are centered on the midpoint of the scale, with the low and high sides of the scale extended to the left and right respectively.

The empirical SI and MI rates are shown together in Figure 4.4. In this plot, the phenomenon of SCALAR DIVERSITY, or variation in by-scale SI rates (Van Tiel et al., 2016), would be shown by variation along the X-axis, which is additionally shown on the top margin with a histogram. A by-item breakdown for both conditions is shown in Appendix C.3, Figures C8 (p. 250) C9 (p. 251). While not one of the goals of this work, we can observe that there appears to be no correlation between SI rates (mean=34.1%, standard deviation=.23) and MI rates (mean=67.8%, standard deviation=.09) and that MI rates are less varied compared to SI rates.

The results of the norming task replicate prior findings on scalar diversity (Van Tiel et al., 2016

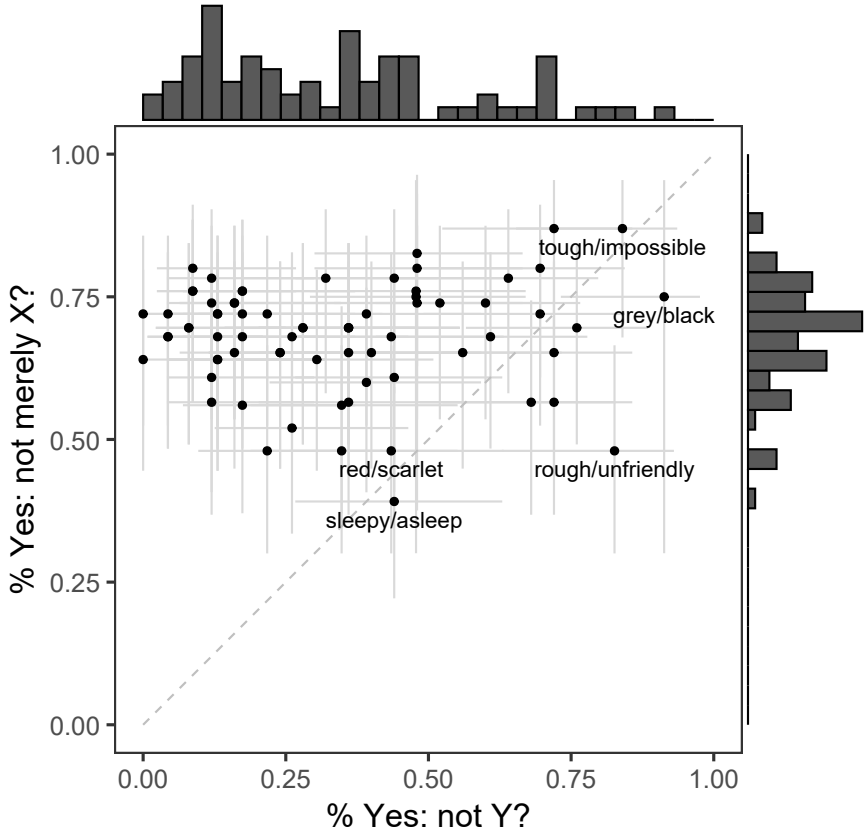


Figure 4.4: Norming task SI and MI rates for both critical conditions. The X-axis reflects SI-rates (*...not cold?*) and the Y-axis reflects MI rates *...not merely cool?*. Error bars indicate 95% Wilson score intervals.

a.o.). Additionally, the indirect question-answer dialogues were generally rated as acceptable. Based on the results, a number of scales were omitted from the remaining experimental materials. Four scales were removed due to a relatively high proportion of low ratings (<*pale*, *white*>, <*sleepy*, *asleep*>, <*cold*, *frosty*>, <*hot*, *scalding*>). Three scales were removed due to *a priori* doubts about whether the SI-enriched interpretation would be available, which was subsequently verified by SI rates close to 0% (<*silly*, *idiotic*>, <*silly*, *ridiculous*>, <*content*, *happy*>). One scale was removed to avoid having three scales with the word *good*, which would complicate the counterbalancing procedure used in the remaining experiments (<*good*, *perfect*>). The remaining 64 scales are used in the remaining experiments.

4.3 Inference Task with Auditory Materials

The goal of this experiment (henceforth Experiment 1) is to determine whether the use of different intonational tunes leads to increased or reduced rates of SI calculation. Based on previous empirical work (de Marneffe & Tonhauser, 2019; Ronai & Göbel, 2024, though cf. Buccola & Goodhue, 2023), use of RFR is predicted to increase the rate of SI computation compared to falls. However, there are a number of potential patterns that may arise from this basic prediction. For example, it may be that any RFR-shaped tune leads to increased SI rates; in ToBI terms, with T* serving as a placeholder for any (rising) pitch accent, the contrast may lie solely in the the edge tone specification: T*L-L% vs T*L-H%. Alternatively, it may be that, more narrowly, the use of either bitonal pitch accent with L-H% edge tones increases SI rates (e.g., Pierrehumbert & Hirschberg, 1990 claim that the use of either bitonal accent invokes a scale).[28] More narrowly still, it might be that it is **specifically** the use of L*+HL-H% that raises SI rates (e.g., following Göbel, 2019).

4.3.1 Materials

The answers (e.g., *The office feels cool*) of the written materials previously described were recorded with myself as the speaker in six intonation conditions, using nuclear tunes crossing one of three pitch accents (H*, L+H*, L*+H) and two edge tone configurations (L-L% and L-H%). The F0 contours were modeled using Generalized Additive Mixed Models (GAMMs) to create standardized contours for pitch resynthesis using PSOLA in Praat (Boersma & Weenink, 2020a). Through the resynthesis procedure, the amount of idiosyncratic phonetic variation by utterance was minimized. Filler items were also recorded and resynthesized to match the resynthesized critical items. These recordings were then spliced with the corresponding question for each item, which were recorded separately by a female native speaker of MAE. The final materials are shown in Figure 4.5; additional details are provided in Appendix C.2.

[28]This raises the additional question of whether the use of a bitonal accent would have the same effect regardless of the edge tone used. Such a pattern would be novel evidence for a strict compositional account of intonational meaning such that the mechanism underlying intonation's role in SI computation is linked specifically to the bitonal accents. However, this is an empirical question and, to foreshadow the results, is not a pattern that arises.

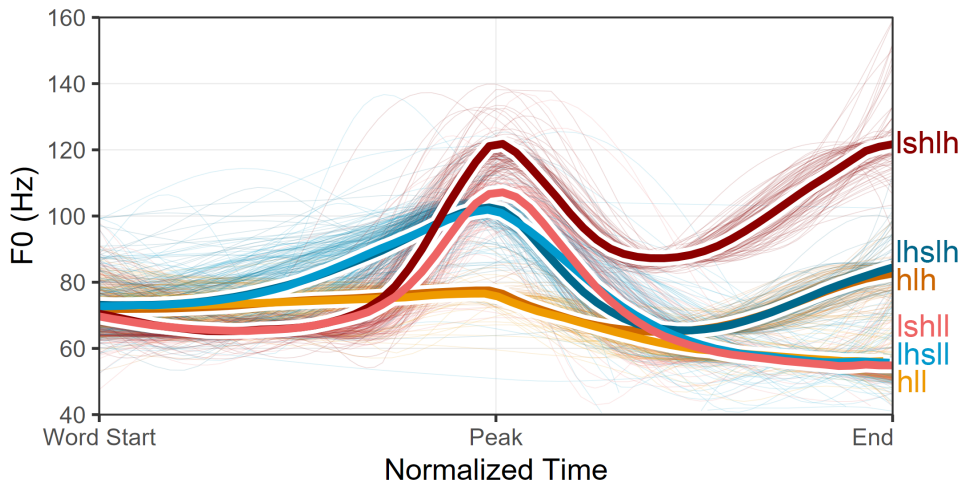


Figure 4.5: Resynthesized materials with superimposed averages. Tune labels use an abridged annotation (L*+HL-H%⇒lshlh). See also the raw recordings in Fig. C2 and the other figures in Appendix C.2.

4.3.2 Procedure

The task procedure is similar to prior implementations of the inference task for SI and scalar diversity (Ronai & Göbel, 2024; Van Tiel et al., 2016). On each trial, participants listen to one of the pre-recorded dialogues (as in 17) in one of six intonation conditions. Participants are then prompted with a Yes/No comprehension question identical to those used in the norming task (which either probe SI or MI computation). Note that unlike the norming task, participants are **not shown the dialogue** in text on the screen—it is only played via audio. Participants listen to 64 critical trials, 36 items from the filler item set, and 72 filler items. Of these fillers, 36 items were adapted from Husband & Ferreira (2016) (described further in the context of priming tasks in Section 4.4.1), separated evenly into 4 blocks. An example of an item from each item set is shown in (17):

(17) a. **Critical**

Question: Did someone leave a window open in the office overnight?

Answer: The office feels cool.

Probe: Would you conclude that the office does not feel cold?

b. **Filler**

Question: Is there an electric car charging station around here?

Answer: We have a gas station.

Probe: Would you conclude that there is no charging station?

c. **HF16-Adapted Item**

Question: Did the museum deliver any good news?

Answer: They thrilled the sculptor.

Probe: Would you conclude that the museum did not thrill the painter?

Unlike the norming experiment, which split the items into conditions probing SI rates and MI rates, all items in this task are presented solely in the SI-probing condition with the exception of pairs of items that happen to share a scalemate (e.g., <*difficult*, *impossible*> and <*tough*, *impossible*>).[29] For these pairs of items, one item is shown in the MI-probing condition used in the norming task (i.e., *...not merely difficult?*) while another is presented in the SI condition (i.e., *...not impossible?*). This way, participants are not asked the same probe question multiple times (for pairs sharing the higher alternative) and also do not hear the same lower alternative (for pairs sharing the lower alternative). Thus, 56 of the 64 items are shown in the SI-probing condition while 8 items are shown in the MI-probing condition.[30] The six intonational tunes are balanced within item set for each participant and the critical items are counterbalanced between participants into 12 lists. Trials are pseudorandomized using a shuffling algorithm that minimizes adjacent trials having the same intonational tune or item set within each block.[31]

[29]To be clear, the SI-probing condition uses the lower alternative in the dialogue and asks a question about the higher alternative. The MI-probing condition uses the higher alternative in the dialogue and asks a question about the lower alternative. So, for a scale like <*cool*, *cold*>, only the dialogue where "The office feels **cool**" is used (with different tunes by participant), and this dialogue is only paired with the SI-probing question "Would you conclude that the office does not feel **cold**." There is no condition that uses the dialogue "The office feels **cold**" paired with the SI-probing question "Would you conclude that the office does not feel **cold**?".

[30]This is a compromise between an experiment that shows no MI-probing questions (maximizing data for SI rates) and an experiment that shows both SI and MI-probing questions evenly. Since the main research question is about the effect of intonation on SI, not MI, the MI data are not directly relevant. On the other hand, by including a limited number of MI trials, we reduce that chance that participants develop a response strategy that is specific to SI and disconnected from a related MI inference task. Note also that the forthcoming lexical decision tasks require both types of items to be shown.

[31]The same algorithm is used to shuffle the trials in the forthcoming lexical decision experiments. The algorithm's efficacy was tested via generating 288,000 unique simulated experiment trial orders, yielding no repeated trial orders while meeting the mentioned criteria for adjacent trials.

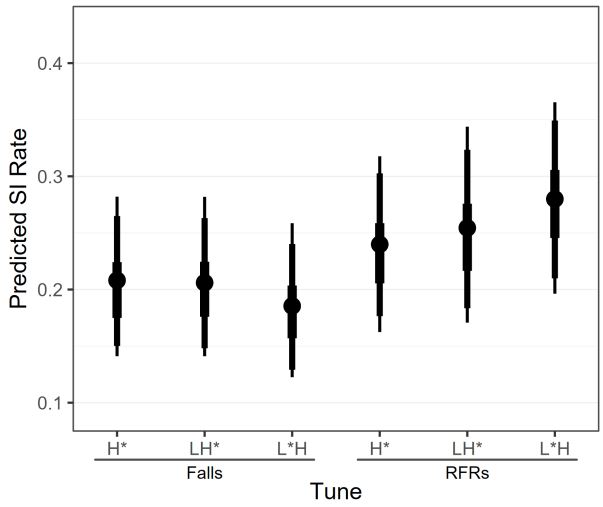


Figure 4.6: Exp. 1 (auditory inference task) model-predicted posterior mean SI rates with 50/89/95% mean highest density intervals.

4.3.3 Results

Participants (n=85) were recruited from Prolific, two of whom were excluded due to self-reporting as having hearing problems. A total of 83 participants (34F, 48M, 1 Other; mean age 38.75) were available for analysis.[32] The task took on average about 30 minutes to complete.

The posterior predicted mean SI rates (model described below) are shown in Figure 4.6. From the figure, it appears that the pitch accents show a numerical cline within the RFR group, with increasing SI rates from H* to L+H* to L*+H. The falling tunes show a decreasing pattern, though notably weaker. The by-item empirical SI rates for each tune are reported in Appendix C.4, Figures C11 and C12.

The likelihood of an SI-enriched interpretation (=Yes response) was modeled using Bayesian logistic mixed effects regression. The model contains a fixed effect of tune (a six-level predictor) with random intercepts and slopes of tune by both participant and item. The contrasts for tune are coded such that this predictor encodes 5 different comparisons: the difference between the RFR and Fall groups, then the differences between L+H* versus H* and L*+H versus H* within each broad tune class; the contrast matrix is shown in Appendix C.4 Table C2. In other words, these

[32]The results of this experiment were previously reported in Sostarics et al. (2025).

contrasts serve to address the difference between the broad tune classes as well as the apparent clines across the pitch accents within each tune class. The model results are shown in Table 4.1.

Term	Estimate	Std.Error	95% CrI	PD
LH*−H*LL	−0.01	0.15	[−0.31, 0.28]	53.24
L*H−H*LL	−0.15	0.16	[−0.46, 0.16]	82.42
RFRs−Falls	0.33	0.09	[0.16, 0.52]	99.95
LH*−H*LH	0.08	0.15	[−0.23, 0.37]	69.17
L*H−H*LH	0.21	0.14	[−0.07, 0.49]	92.73

Table 4.1: Exp. 1 (auditory inference task) logistic regression model results for by-tune SI rates. Estimates are given on the log-odds scale.

The statistical model shows a main effect of tune class: RFR-shaped tunes have higher SI rates than Falls ($\hat{\beta} = 0.33, CrI = [0.16, 0.52]$). Within each tune class, there is not much evidence that the bitonal accents behave differently from H* for either the Falls ($\hat{\beta} = -0.15, CrI = [-0.46, 0.16]$) or the RFR-shaped tunes ($\hat{\beta} = 0.21, CrI = [-0.07, 0.49]$). At best, the probability of direction for the L*+H versus H* comparison is 92.73%, suggesting weak evidence for higher SI rates for L*+HL-H% compared to the monotonal H*L-H%—this point will be revisited in the general discussion.

4.4 Cross-modal Lexical Decision Tasks

Whereas the auditory inference task of Exp. 1 addressed the question of differences in offline interpretation given different intonational tunes in the context of SI, the cross-modal lexical decision experiments, discussed next, address the potential effects of different tunes on online lexical processing. Based on the review of previous accounts of RFR, we know that RFR is related in some way with alternatives—on some accounts, higher alternatives specifically. But it remains an open question whether SI leads to the activation of alternatives in a similar way as seen with the processing of focus alternatives (Gotzner & Spalek, 2019).

4.4.1 Adapting Husband and Ferreira's Materials

The items from Husband & Ferreira (2016) are adapted to the question-answer format described in the previous section to serve as a set of filler items with real-word TARGETS. Using the prime *sculptor* as an example, recall that these items contain three target conditions: a CONTRASTIVE condition (*painter*), a NON-CONTRASTIVE condition (*statue*), and a semantically UNRELATED condition (*register*). So, in addition to serving as real-word fillers for this experiment, these items also allow us to make the comparison between contrastive alternatives and scalemates (i.e., the critical items). An example of an adapted item is shown in (18); this item set will continue to be referred to as the **HF16-adapted items**.

(18) (*Original item from HF-16: The museum thrilled the* ***sculptor*** *when they called about his work.*)

Alice: Did the museum deliver any good news?

Bob: The museum thrilled the **sculptor**.

Targets: CONTRASTIVE*: painter,* NON-CONTRASTIVE*: statue,* UNRELATED*: register*

4.4.2 Procedure

Participants were instructed that they will be listening to dialogues and judging whether a string of letters that appears on the screen is a word or not a word of English. Each trial begins with a fixation cross appearing on the screen for 1 second. Participants then listen to a question/answer dialogue (the same materials from the previous experiment) which ends in the auditory PRIME. After a delay (known in this literature as the *stimulus onset asynchrony*, or SOA), the visual target appears on the screen; this work follows Husband & Ferreira (2016) in using two different SOAs across experiments: Exp. 2, 2b, and 4 use a long SOA of 750ms (the SOA at which Husband & Ferreira, 2016 found differences based on prosody) while Exp. 3 uses a short SOA of 0ms (to probe an earlier timepoint in processing). Using a buttonbox, participants are tasked with judging whether the string of letters that appears on the screen is a word or not a word of English by

pressing one of two buttons (mapped to either Yes or No), after which the next trial begins. The mapping of Yes/No to the left and right buttons is randomized by participant.

Prior to the main task, participants were familiarized with the task. Participants see 24 practice lexical decision trials with half non-word targets and half real-word targets and an inter-stimulus interval of 750ms after participants respond; there is no audio component to these practice trials. Participants are given feedback on accuracy and speed (if slower than 6 seconds) with no opportunity to retry. Following this, participants see 3 more practice trials **with** audio that have the same structure as the critical trials; participants no longer receive feedback on these trials.

Each participant sees 186 trials: 64 critical items split evenly[33] into twelve conditions (HIGHERTARGET or LOWERTARGET, with one of six intonational tunes), 61 real-word items from the HF16-adapted item set split into three conditions (based on the target, either CONTRASTIVE, NONCONTRASTIVE, or UNRELATED), and 61 non-word items (the filler item set described in Section 4.2.1). To make this maximally explicit: Yhe HIGHERTARGET condition is the one where participants **hear *cool* and see *cold*** and the LOWERTARGET condition is the one where participants **hear *cold* and see *cool***. The HF16-adapted items and filler items are split evenly (*modulo* 1) into the six intonational tunes such that specific tunes aren't exclusively or disproportionately associated with the critical item set. The 186 trials are split into 6 blocks of 30 to 32 trials using the shuffling algorithm previously described. As in the previous experiments, the critical items are counterbalanced into 12 lists while the HF16-adapted items are counterbalanced into 3 lists.[34] Participants had a mandatory 10 second break between each block, but are allowed to take longer if desired. RT is measured as the time between when the word appears on the screen to when the participant presses a button.

The experiment is administered in-person in a sound-attenuated booth to minimize environmental distractions (a comparison with an online implementation is given in Section 4.4.3.5). The

[33]Because 64 is not evenly divisible by 12, with any given critical item list, there are four items left over. Thus, four conditions are instantiated by six items while the remaining eight items are instantiated by five items. No item is repeated. The same is true for the HF16-adapted items (*modulo* one instead of four; i.e., one condition per list has 21 items while the others have 20).

[34]The two sets of lists were manually rotated such that each critical item list appears with each HF16-adapted item list with the goal of keeping the assignments balanced. This was later automated for future experiments.

experiment is implemented in PsychoPy (Peirce, 2007), which synchronizes the onset of stimuli to the refresh rate of the monitor, which allows stimulus timings to be accurately measured. A 165Hz refresh rate monitor was used to minimize latency between the offset of the audio stimulus and the onset of the visual target. A Cedrus RB-740 button box was used to minimize latency between when a response is made and when it is registered by the hardware.

4.4.3 Long SOA (750ms) Results

Undergraduate students at Northwestern University participated for course credit (n=104). None of these participants participated in Exp. 1, which recruited participants from prolific. Participants were excluded if they self-reported as not being native speakers of MAE (n=37) or if they displayed overly dispersed RT distributions indicative of inattention during the task (n=3). In addition, one participant was excluded due to a reported neuropathy that impacted movement of the hand. A total of 63 participants remained for further analysis (34F, 26M, 3 Other, mean age 19.8). The experiment took approximately 20-35 minutes for participants to complete.

Participant RTs when correctly responding YES (total accuracy 98.6%) were modeled using a Bayesian lognormal distributional model to address the three questions laid out in Section 4.1.4.1. RTs faster than 200ms or slower than 1500ms were discarded, resulting in a data loss of 0.49%. As a reminder, the first question is whether scalar alternatives behave differently from focus/contrastive alternatives (i.e., on the basis of lexical relationship, averaging over intonation). Incidentally, addressing this question will also assess, as a sanity check, whether there are effects of semantic priming (where semantically related words should be faster than unrelated words). The second question is whether one or more RFR-shaped tunes offer a processing advantage for the HIGHERTARGET condition (i.e., is there a simple effect of tune in this condition). The last question is whether there is an asymmetry in the behavior of one or more RFR-shaped tunes when looking at the LOWERTARGET condition (i.e., is there an interaction between tune and target condition).[35] This third point is crucial: if RFR is specifically related to **higher** alternatives, then

[35]Note that although all tunes are used in the HF16-adapted item set, each item is only recorded with a single tune (c.f. every critical item is recorded with every tune). As a result, we cannot probe further questions such as whether the

RTs should differ for the HIGHERTARGET condition compared to the LOWERTARGET when RFR is used.[36]

The model is parameterized in such a way to make answering these three specific questions straightforward, but has additional complexities that are needed for the actual modeling but not for inferential purposes. Only the statistics relevant for the three main questions described above are reported in the main text; additional details are available in Appendix C.5. The main predictors of interest are Condition, which is Helmert coded to encode nested comparisons of each target condition.[37] The intercept is set at the HIGHERTARGET condition mean. Tune is sum coded (± 1) to encode deviations from the HIGHERTARGET condition mean. An interaction between Tune and Condition is included, but only the interaction terms related to the LOWERTARGET/HIGHERTARGET conditions will be reported in the main text. The model also controls for effects of log word frequency (Balota et al., 2007) of the target word, length of the target word (as the number of letters), and experimental block—controls are treated as continuous and centered on their means. The random effects structure includes random intercepts by participant and item and random slopes of Tune, Condition, and their interaction by participant and by item. The discrimination parameter in this model is parameterized only with random intercepts by participant and item; this parameter can be considered to account for differences in RT dispersion/variability for individual participants and items. The model results will be presented incrementally across multiple tables (rather than all at once) as they become relevant for addressing each question. The full model output is available in Table C4 in C.5.

present results replicate those reported in Husband & Ferreira (2016) or how the RFR-shaped tunes interact with the non-scalar alternatives, because intonation does not vary within item and each tune is only instantiated by 10 HF16-adapted items. The inclusion of the HF16-adapted items is solely to address complementary questions related to the lexical properties of the target/prime pairs while also providing a more representative dataset with which to control for effects of word length/frequency when looking at the critical items.

[36]To make the importance of the LOWERTARGET condition maximally explicit: If we find facilitation in the HIGHERTARGET condition when RFR is used, then without the LOWERTARGET condition we would not know if the facilitation is due to a specific connection to **higher** alternatives, or whether facilitation arises due to a broader connection to **alternatives** more generally. Further, if facilitation were equated with increased SI computation, then similarly we would not be able to conclude whether the facilitation arises from SI computation or due to accessing alternatives.

[37]Condition has 5 levels, and so the 4 comparisons test (1) whether higher alternatives differ from lower alternatives; (2) whether scalemates differ from non-scalemate contrastive focus alternatives; (3) whether contrastive semantic associates differ from non-contrastive associates; and (4) whether semantically-related words differ from semantically unrelated words. These correspond to each branching point in Figure 4.2.

4.4.3.1 Are Scalemates Different from Contrastive Alternatives?

To address whether whether scalemates behave any differently from non-scalar contrastive alternatives, we need to look at the relationship between the lexical items alone. The intonation with which the response is uttered is not at issue. Hence, Figure 4.7 shows the model-predicted RTs for each target condition, which is aggregated across all intonation conditions. The relevant model parameters are the fixed effects of the Condition predictor, which is shown with the intercept in Table 4.2

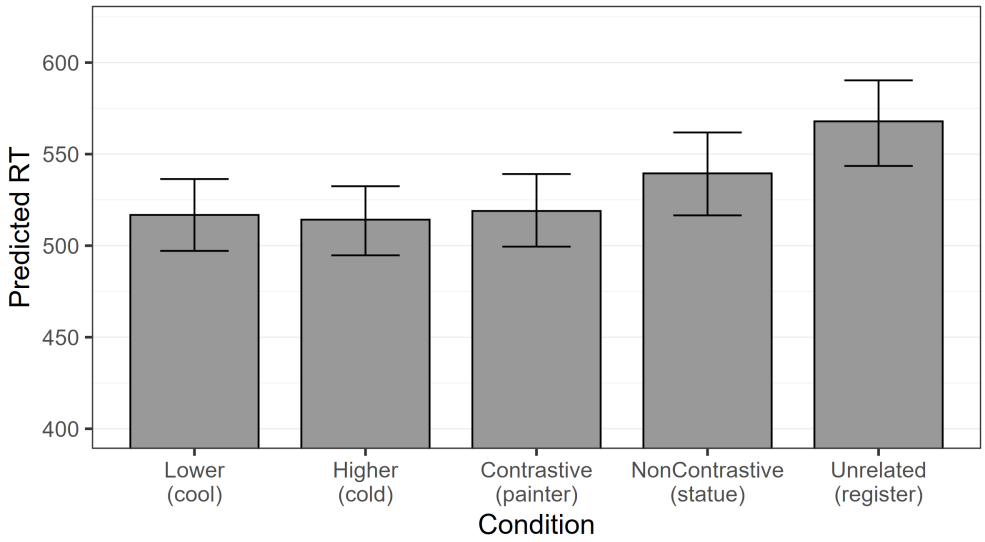


Figure 4.7: Exp. 2 (long SOA lexical decision) model-predicted RTs with 95% mean highest density intervals. RTs are marginalized over target word length and frequency and experimental block.

Term	**Estimate**	**Std.Error**	**95% CrI**	**PD**
HigherTargetMean	6.226	0.019	[6.189, 6.262]	100.00
HigherTarget−LowerTarget	−0.005	0.012	[−0.028, 0.018]	66.69
NonScalar−Scalar	0.007	0.012	[−0.017, 0.030]	71.35
NonContrastive−Contrastive	0.043	0.014	[0.016, 0.070]	99.85
Unrelated−Related	0.084	0.013	[0.059, 0.108]	100.00

Table 4.2: Exp. 2 (long SOA lexical decision) fixed effects of condition. The intercept corresponds to the mean of the HIGHERTARGET condition.

The model results show that all targets that are semantically related to their primes are credibly faster than semantically unrelated words ($\hat{\beta} = 0.084, CrI = [0.059, 0.108], PD = 100.00$).

Breaking down the semantically related words into non-contrastive versus contrastive associates, the model shows that contrastive associates are faster than non-contrastive associates ($\hat{\beta} = 0.043, CrI = [0.016, 0.070], PD = 99.85$). Breaking down the contrastive associates, the model does not not show evidence that the scalemates are credibly faster than the non-scalar contrastive alternatives ($\hat{\beta} = 0.007, CrI = [-0.017, 0.030], PD = 71.35$). Finally, breaking down the scalemate conditions, the model does not show evidence that the HIGHERTARGET and LOWERTARGET conditions are credibly different from one another ($\hat{\beta} = -0.005, CrI = [-0.028, 0.018], PD = 66.69$). Taken together, these results show that there is facilitation due to semantic priming (semantically related words are faster than unrelated words) and additional facilitation when the target word is able to serve as a contrastive focus alternative. However, the scalemate relation does not afford any additional facilitation overall.[38]

4.4.3.2 Do RFR-Shaped Tunes Yield a Processing Benefit for Higher Alternatives?

The next question is whether the different intonational tunes modulate RT within the HIGHERTARGET condition specifically. Visually speaking in terms of the bar plot in Figure 4.7, the goal is to break up the individual Lower and Higher bars into the six tunes and see whether one or more tunes push the bar higher or lower. A tune that pushes the bar lower would be reflective of faster RTs in that condition when using that tune. The results will be described in terms of percent change ($\%\Delta$) from the condition means.[39] Importantly, because the condition results previously reported show that both scalemate conditions are lower than the semantically unrelated condition, the effects reported here will describe the **degree** of facilitation (i.e., more or less facilitation as opposed to facilitation vs inhibition).

Because RFR is described as dealing with **higher** alternatives, the primary condition of inter-

[38]It should be emphasized again that the results presented in Figure 4.7 aggregate over all tunes. In the context of the HF16-adapted items, this means that the 10 items in the "neutral prosody" condition and the 10 items in the "focus prosody" condition are averaged together here. The question of whether the findings of Husband & Ferreira (2016) are directly replicated here is not answerable with this dataset because tune does not vary within each HF16-adapted item.

[39]Because the models here use a lognormal likelihood (i.e., they work with logRTs and not RTs), coefficients ($\hat{\beta}$) can be interpreted in terms of percent change via the transformation $100(e^{\hat{\beta}} - 1)$. Here, percent change is equal to the proportional speedup or slowdown of RT (e.g., 2% faster or slower).

est is the HIGHERTARGET condition. As mentioned previously, the prediction here is that one (or maybe more) RFR-shaped tune(s) will lead to additional facilitation in this condition.[40] The relevant model parameters are the fixed effects of each tune, which encode deviations from the HIGHERTARGET mean; these effects are shown with the intercept in Table 4.3. The posterior-predicted percent change distributions are shown in Figure 4.8 (note the correspondence between the values in the right panel and the $\%\Delta$ values in Table 4.3).

Term	**Estimate**	$\%\Delta$	**Std.Error**	**95% CrI**	**PD**
HigherTargetMean	6.226		0.019	[6.189, 6.262]	100.00
H*LH	−0.020	−1.95	0.008	[−0.035, −0.004]	99.30
LH*LH	0.005	0.46	0.008	[−0.011, 0.021]	71.56
LH*LL	−0.010	−1.02	0.008	[−0.026, 0.006]	89.83
L*HLH	0.019	1.94	0.008	[0.003, 0.035]	99.14
L*HLL	0.006	0.62	0.008	[−0.010, 0.022]	77.49

Table 4.3: Exp. 2 (long SOA lexical decision) fixed effects of tune with the intercept, which corresponds to the mean of the HIGHERTARGET condition. Estimates are shown on the natural log scale with associated percent change values.

From Figure 4.8 we can see that, generally, there is very little variation in RT among tunes in the LOWERTARGET condition. This pattern is in line with the prediction that variation in processing when RFR is used should be related to the processing of the higher alternative, not the lower alternative. More broadly, this pattern indicates that intonation has an effect beyond mere semantic relatedness of the prime and target words or whether the target word can serve as a contrastive alternative to the prime (as shown in Figure 4.7).[41] Looking next at the HIGHERTARGET condition, H*L-H% is credibly faster than the condition average ($\hat{\beta} = -0.020, \%\Delta = -1.95\%, CrI = [-0.035, -0.004], PD = 99.30$) and L*+HL-H% is credibly slower than the condition average.

[40]To temper expectations here, the effect sizes are not expected to be dramatically large, particularly in comparison to the condition comparisons shown in the previous section (Table 4.2). The finding of Husband & Ferreira (2016, p. 226), where non-contrastive associates were slower than contrastive associates with focus prosody, was a percent change of only 3.4%; the difference between the prosody conditions for any given target in that experiment was between 0.08% and 2.3%.

[41]To reiterate and make this maximally explicit: any effect of semantic relatedness between *cold* and *cool* is expected to exist regardless of the intonation with which the prime is uttered. To my knowledge, no account would predict *cold* uttered with RFR to be more semantically similar to *cool* than *cold* uttered with any other intonational tune. Therefore, any effect of semantic relatedness is already reflected in Figure 4.7.

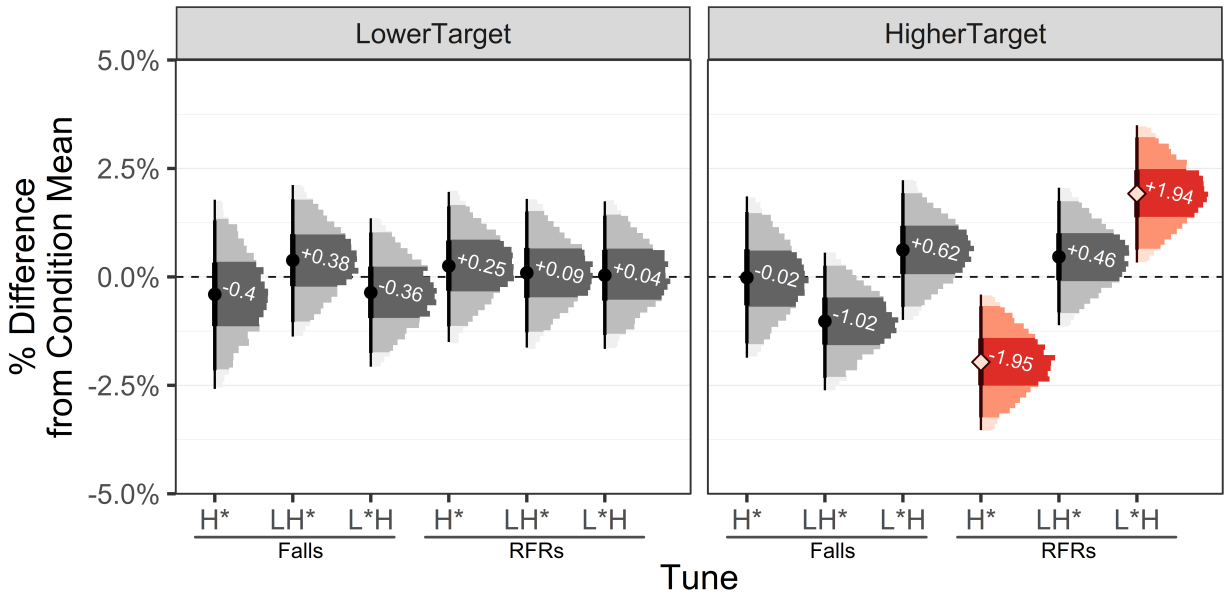


Figure 4.8: Exp. 2 (long SOA lexical decision) model-predicted posterior distribution of RT percent changes with 50%, 89%, and 95% mean highest density intervals. Tunes that are not credibly different from the condition means are shown in gray (means=black circles) while tunes that are credibly different are shown in red (light diamonds).

($\hat{\beta} = 0.019, \%\Delta = 1.94\%, CrI = [0.003, 0.035], PD = 99.14$) L+H*L-H% lies somewhere between them, near the condition mean ($\hat{\beta} = 0.005, \%\Delta = 0.46\%, CrI = [-0.011, 0.021], PD = 71.56$). None of the three falling tunes are credibly different from the condition average.

4.4.3.3 Is there an Asymmetry for RFR-Shaped Tunes?

The last question to be addressed is whether there is an asymmetry in the processing profile of the RFR-shaped tunes when comparing across the LOWERTARGET and HIGHERTARGET conditions. That is, if RFR is related to higher alternatives specifically, and we expect this to lead to facilitation, then it should not simultaneously lead to facilitation of the lower alternative. Such a pattern would suggest that RFR does not have a connection to higher alternatives specifically but is instead related to scalar alternatives more broadly. These comparisons are encoded by the interaction terms between Tune and Condition in the model, shown in Table 4.4.

Based on the statistical model, we find evidence of an asymmetry for both H*L-H% ($\hat{\beta} = -0.022, CrI = [-0.044, 0.000], PD = 97.47$) as well as L*+HL-H% ($\hat{\beta} = 0.019, CrI =$

Term	Estimate	%Δ	Std.Error	95% CrI		PD
H*LH	−0.022	−2.19	0.011	[−0.044,	0.000]	97.47
LH*LH	0.004	0.37	0.011	[−0.018,	0.026]	62.60
LH*LL	−0.014	−1.39	0.011	[−0.036,	0.009]	89.28
L*HLH	0.019	1.90	0.011	[−0.003,	0.041]	95.42
L*HLH	0.010	0.99	0.011	[−0.012,	0.032]	80.63

Table 4.4: Exp. 2 (long SOA lexical decision) interaction terms between Tune and Condition, which correspond to the total difference between the HIGHERTARGET and LOWERTARGET conditions for each tune. See Appendix C.5 for more information. Estimates are shown on the natural log scale with associated percent change values.

$[-0.003, 0.041], PD = 95.42$) based on the probability of direction for each interaction.[42] No other tunes show an asymmetry. Importantly this result is in line with the hypothesis that RFR is related specifically to higher alternatives: while H*L-H% is slightly faster and L*+HL-H% is slightly slower, these patterns are not present in the LowerTarget condition.

4.4.3.4 Interim Discussion

The results reported so far provide the bulk of the evidence for our research questions, so it is worth discussing the data from Experiments 1 and 2 in relation to the main research questions. To foreshadow the remaining experiments: the results are more variable and do not relate as cleanly to the stated research questions.

As a reminder, this work asks three main questions. The first question is whether RFR-shaped tunes that differ in the specification of the pitch accent behave similarly or differently from one another in offline interpretation (Exp. 1) and/or online processing (Exp. 2). The pattern of results between the two experiments is somewhat counterintuitive. As a broad class of tunes, the

[42]These results may seem obvious based on Figure 4.8, so to better appreciate and understand them, take for example the distributions for L+H*L-L% (second from the left in each facet). When looking across facets, we can see that there is overlap between these two distributions despite the means going in opposite directions (+0.38 versus -1.02, for a total distance of ≈1.40). Because of the uncertainty in the estimates for this tune in each condition, we do not find credible evidence that they're behaving differently from one another. In contrast, when looking at L*+HL-H% (far right in each facet), there is also some overlap in the distributions between the two facets/conditions. Yet unlike L+H*L-L%, we do identify a difference between these two—there's some overlap, but there is enough differentiation for the probability of direction for the difference between them to be greater than 95%. So, the asymmetry question addressed here involves explicitly testing whether the two distributions are credibly different from one another.

RFR-shaped tunes had overall higher SI rates than the Falls did, with some potentially graded distinctions within this broad class such that RFR with L*+H had the highest SI rates. Yet, in online processing, we find that H*L-H% leads to additional facilitation of the higher alternative beyond what would be expected based on semantic association alone, while L*+HL-H% instead leads to less facilitation of the higher alternative. Here, the counterintuitive pattern that emerges is that RFR-shaped tunes behave similarly in offline interpretation yet differently in online processing. Moreover, the pattern seems to suggest a negative relationship between likelihood of SI and priming. This pattern runs counter to what would be expected from the focus alternative literature, which overwhelmingly finds a facilitatory effect of focus despite the negation of the probed alternative (Gotzner & Spalek, 2019). In summary, we find a counterintuitive pattern that RFR-shaped tunes seem to behave both differently and similarly, depending on the type of task used. These findings will be revisited in the general discussion to propose an account to unify these seemingly contradictory results in terms of within-category variation.

The second research question is whether scalar alternatives behave differently from focus alternatives in processing on the basis of the lexical relation between the prime and target word. Recall that a pair of lexical items like <*cool*, *cold*> comprise a lexical scale and so are related via asymmetric entailment—they are taken to be scalar alternatives to one another. Yet, these items can also be contrastive focus alternatives to one another. Despite ample work on the processing of focus alternatives, it is not clear whether scalar alternatives would enjoy a processing advantage above and beyond their status of being able to be focus alternatives to one another. For instance, the alternative generating process may be the same for scalar and focus alternatives (Fox & Katzir, 2011). The processing results of Exp. 2 suggest that scalar alternatives do not have an additional processing advantage over merely being able to be contrastive alternatives. The comparison between our scalar and contrastive conditions, however, comes with the caveat that intonational tune was not manipulated within the contrastive target condition. Consequently, these results speak only to whether the target is **able to be** a contrastive alternative to the prime.

The results addressing the third question of whether there is an asymmetry in the processing

correlate of RFR such that it **specifically** targets the higher alternative, and not the lower alternative, are relatively more straightforward. The results showed evidence for such an asymmetry for both H*L-H% and L*+HL-H%; while there was no effect of L+H*L-H%, nor an asymmetry, it should be noted that its distribution appears between the two RFR-shaped tunes in much the same way as the apparent cline in SI rates from the auditory inference task (Exp. 1) results. Setting this aside, it does appear that there is an asymmetry for RFR tunes, though an account for the counter-intuitive direction and apparent clines in the data will be revisited in the general discussion (§4.5). Additionally, the asymmetry that RFR selectively targets the higher alternative, and not the lower alternative, suggests that the scalar relationship is nonetheless important—in contrast to the finding that scalar conditions were no different from the contrastive condition. Overall this is a welcome result: the assumption that RFR is related to higher alternatives is supported by the processing results.

The results presented thus far are also particularly meaningful in relation to the findings from De Carvalho et al. (2016). In their study, they argued that under a Neo-Gricean (NG) theory, stronger scalar terms are necessary to the interpretation of weaker terms due to SI being automatic, but the reverse (stronger terms needing their weaker terms for interpretation) is not necessary. Accordingly, they report an asymmetry in priming such that weaker terms prime their stronger terms more than the reverse. They took these results as evidence for NG accounts of SI and more broadly as evidence for the psychological reality of lexical scales, as otherwise the asymmetry would not arise. The results presented in this work (specifically, the results shown in Fig. 4.7) do not replicate their finding: Based on the condition-level comparisons, we did not see an inherent asymmetry between the HIGHERTARGET and LOWERTARGET conditions. In other words, *cool* does not prime *cold* more than the reverse.[43] Therefore, the results presented here are against the claim that the stronger terms are necessary to the interpretation of weaker terms. However, the by-tune results (showing an asymmetry in the behavior of RFR between the HIGHERTARGET and

[43]In fact, when not controlling for word length and frequency, the HIGHERTARGET condition generally has slower, not faster, RTs compared to the LOWERTARGET condition—largely because the higher alternatives are typically longer and less frequent.

LOWERTARGET conditions) nonetheless point towards a distinction between stronger and higher scalemates, which is in line with the broader claim from De Carvalho et al. (2016) that lexical scales are psychologically real.

4.4.3.5 Online Version of the Experiment

Building on the results of Exp. 2, an online version of the lexical decision experiment was conducted to determine whether the effects of intonation are robust even in a less controlled environment. The practical question addressed here is whether it is feasible to obtain more data on SI and Lexical Decision without going to the great lengths to control the hardware and environment for the participant required for the in-person experiment.

The implementation of Exp. 2 was automatically transpiled from PsychoPy to PsychoJS and hosted on an independent Google Firebase web server, which also handled the condition counterbalancing and assignment. For this experiment, 60 participants were recruited from Prolific. One participant was excluded due to self-reporting that they did not grow up in the United States, leaving 59 participants for further analysis. The data was analyzed using the same model described for the previous experiment.

The results are rather straightforward: The data collected from online participants is noisier and there are no effects of intonation in any condition. However, the effects of semantic priming ($\hat{\beta} = 0.053, CrI = [0.026, 0.080], PD = 99.98$) and non-contrastive versus contrastive alternatives ($\hat{\beta} = 0.025, CrI = [0.000, 0.050], PD = 97.63$) still persist. This result suggests that cross-modal lexical decision tasks looking at intonational factors are best left to controlled environments, and so web-based implementations are thus abandoned for subsequent experiments. Further discussion of these results is assigned to Appendix C.5.1, but it should be emphasized here that this experiment was not fully in vain. Specifically, the results provide a welcome methodological contribution regarding the feasibility of cross-modal lexical decision tasks in uncontrolled environments.

4.4.4 Short SOA (0ms) Results

Experiment 2 investigated the processing profile of different intonational tunes at a relatively later timepoint in processing (an SOA of 750ms). Following Husband & Ferreira (2016), Experiment 3 also investigates the early processing profile of the same tunes using a 0ms SOA. The in-person setup is the same as Experiment 2.

4.4.4.1 Results

Undergraduate students at Northwestern University participated for course credit (n=84). Participants were excluded if they self-reported as not being naive MAE speakers (n=15) or had hearing or reading deficits (n=6). A total of 63 participants were included for further analysis (37F, 25M, 1 Other; mean age 19.96). The model is the same as in the previous experiment with the exception of placing the intercept at the mean of the LOWERTARGET condition rather than the HIGHERTARGET condition.[44] The condition results are shown in Figure 4.9, with the fixed effects of condition reported in Table 4.5.

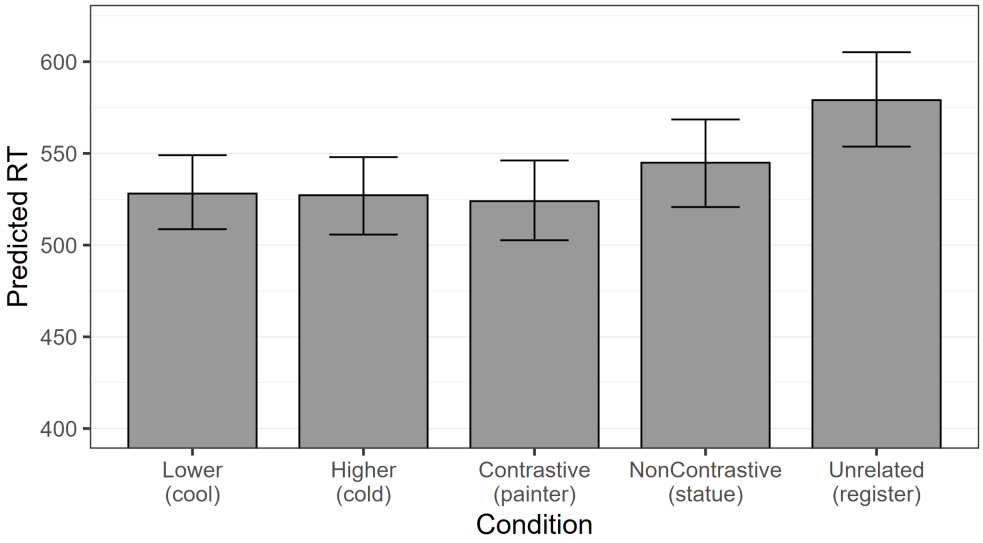


Figure 4.9: Exp. 3 (short SOA lexical decision) posterior-predicted mean RTs for each target condition with 95% highest density intervals.

[44]The reason should be evident upon seeing Figure 4.10; this decision only changes the interpretation of the the coefficients involving the Tune predictor.

Term	Estimate	Std.Error	95% CrI	PD
LowerTargetMean	6.252	0.020	[6.214, 6.290]	100.00
HigherTarget−LowerTarget	−0.002	0.012	[−0.025, 0.022]	55.78
NonScalar−Scalar	−0.007	0.013	[−0.033, 0.018]	70.47
NonContrastive−Contrastive	0.035	0.013	[0.009, 0.060]	99.55
Unrelated−Related	0.087	0.015	[0.057, 0.115]	100.00

Table 4.5: Exp. 3 (short SOA lexical decision) fixed effects of condition with the intercept, which correspond to the mean of the LOWERTARGET condition.

Similar to the long SOA experiment (Exp. 2), the model shows that semantically unrelated words are slower than semantically related words ($\hat{\beta} = 0.087, CrI = [0.057, 0.115], PD = 100.00$) and that non-contrastive associates are slower than contrastive associates ($\hat{\beta} = 0.035, CrI = [0.009, 0.060], PD = 99.55$). Again, there is no difference betwen scalar versus non-scalar contrastive alternatives ($\hat{\beta} = -0.007, CrI = [-0.033, 0.018], PD = 70.47$) nor between higher versus lower scalemates ($\hat{\beta} = -0.002, CrI = [-0.025, 0.022], PD = 55.78$).

The effects of each tune for the two scalar target conditions are shown in Figure 4.10. Unlike the previous experiment, the variation in by-tune behavior appears in the LOWERTARGET condition rather than the HIGHERTARGET condition—hence the reason for switching the intercept level. The fixed effects for each tune thus represent deviations from the LOWERTARGET condition mean rather than the HIGHERTARGET condition mean. The fixed effects for each tune are reported in Table 4.6. The statistical model shows an effect of L*+HL-H% such that is faster in the LOWERTARGET condition ($\hat{\beta} = -0.019, \%\Delta = -1.86\%, CrI = [-0.035, -0.003], PD = 99.10$) and an effect of H*L-L% such that it is slower in the LOWERTARGET condition ($\hat{\beta} = 0.019, \%\Delta = 1.91\%, CrI = [0.000, 0.038], PD = 97.70$).[45]

Regarding asymmetries between the two target conditions, reflected by the interactions in the model, there is an asymmetry for L*+HL-H% such that it is credibly faster in the LOWERTARGET condition compared to the HIGHERTARGET condition ($\hat{\beta} = 0.028, CrI = [0.006, 0.050], PD = 99.31$). There is no evidence for an asymmetry with the other tunes, including H*L-L% ($\hat{\beta} =$

[45]Because H*L-L% is used as the reference level in the model, the "effect" and all associated values are computed manually from the posterior distribution.

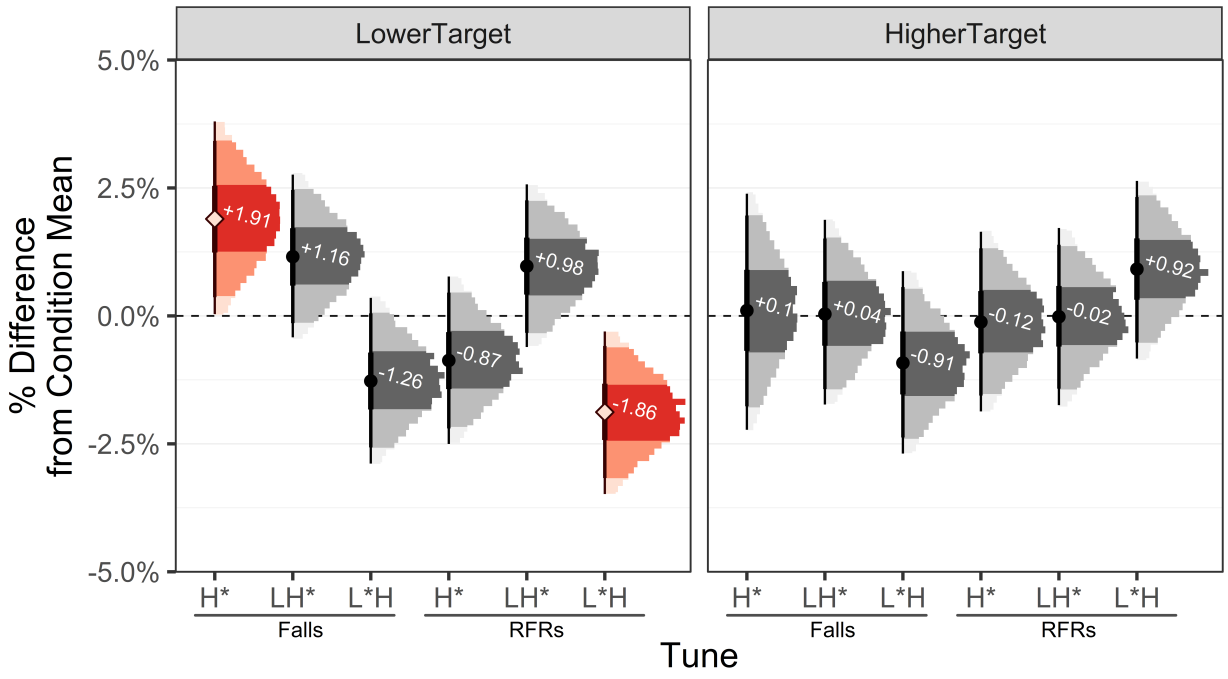


Figure 4.10: Exp. 3 (short SOA lexical decision) model-predicted posterior distribution of RT percent changes with 50/89/95% mean highest density intervals. Tunes that are not credibly different from the condition means are shown in gray (means=black circles) while tunes that are credibly different are shown in red (means=light diamonds).

$0.018, \%\Delta = 1.81\%, CrI = [-0.010, 0.045], PD = 90.00$). The full set of interaction terms are reported in Table 4.7. The remaining model statistics are reported in Table C6 in Appendix C.6.

4.4.5 Dual Task

The experiments presented so far have shown that RFR-shaped tunes yield higher rates of SI computation. However, the inference task explicitly foregrounds SI via repeatedly asking questions of the form *Would you conclude that [something is not _]?* while performance on the lexical decision task is completely orthogonal to SI computation. In other words, one does not need to compute SI or really do any pragmatic reasoning to judge whether a string of letters on the screen is a word or not a word. As a result, we do not know whether the observed facilitation of the higher alternative is the direct consequence of SI computation—it may or may not be. In this last experiment, a dual task approach is used with the goal of relating lexical decision performance to whether or not par-

Term	Estimate	%Δ	Std.Error	95% CrI	PD
LowerTargetMean	6.252		0.020	[6.214, 6.290]	100.00
H*LH	−0.009	−0.87	0.008	[−0.025, 0.008]	85.45
LH*LH	0.010	0.98	0.008	[−0.006, 0.026]	88.04
LH*LL	0.012	1.16	0.008	[−0.004, 0.028]	92.14
L*HLH	−0.019	−1.86	0.008	[−0.035, −0.003]	99.10
L*HLL	−0.013	−1.26	0.008	[−0.029, 0.004]	93.76

Table 4.6: Fixed effects of tune with the intercept, which corresponds to the mean of the LOWERTARGET condition. Estimates are shown on the natural log scale with associated percent change values.

Term	Estimate	%Δ	Std.Error	95% CrI	PD
H*LH	0.008	0.75	0.012	[−0.015, 0.031]	74.25
LH*LH	−0.010	−0.98	0.011	[−0.032, 0.012]	80.55
LH*LL	−0.011	−1.12	0.012	[−0.034, 0.011]	83.01
L*HLH	0.028	2.83	0.011	[0.006, 0.050]	99.31
L*HLL	0.004	0.36	0.012	[−0.019, 0.026]	62.05

Table 4.7: Exp. 3 (short SOA) interaction terms between Tune and Condition, which correspond to the total difference between the LOWERTARGET and HIGHERTARGET conditions for each tune. Estimates are shown on the natural log scale with associated percent change values.

ticipants compute the SI-enriched interpretation based on their response to an SI-probing question. This task is thus a combination of the cross-modal lexical decision task and the inference task.

4.4.5.1 Task Setup and Procedure

For the main task, participants first listen to an auditory question-answer dialogue ending in a prime word such as *cool*. Then, after a 750ms SOA, a target word such as *cold* appears on the screen and participants judge whether Yes, it is a word, or No, it is not a word. Immediately afterwards (i.e., with no repetition of the auditory dialogue), participants are asked a question as in the auditory inference task (Exp. 1) such as *Would you conclude that the office is not cold?*. Note that the dialogue is only presented auditorily and is not written out on the screen (as in Exp. 1). Participants respond either Yes or No using the same response mapping as with the lexical decision task.

The familiarization phase is expanded from two steps to three steps. Like in the lexical decision task, participants do 24 lexical decision trials with no audio and are given feedback on whether they were incorrect or too slow. Then, participants do two lexical decision trials with audio. Finally, participants do both task components on the same trial after listening to an auditorily presented dialogue: judge if a string is a word or not a word, then answer a question that appears after making the lexical decision. Participants do three of these dual-task practice trials (the first two are the same trials as the with-audio lexical decision practice trials). Participants were told that they should read the questions carefully and that they can take as much time as needed to answer, but that the lexical decision portion should be done quickly.

Undergraduate students at Northwestern University participated for course credit (n=82), none of whom participated in any of the previous experiments. Participants were excluded if they self-reported that they were not native MAE speakers (n=5) or self-reported vision or hearing difficulties (n=3), leaving 74 participants available for analysis (39F, 33M, 2 Other, mean age 20.3).

4.4.5.2 Inference Task Results

SI responses that were faster than the majority of the participant's lexical decision responses were excluded, resulting in a data loss of 4.83%. The model predicted SI rates are shown in Figure 4.11.

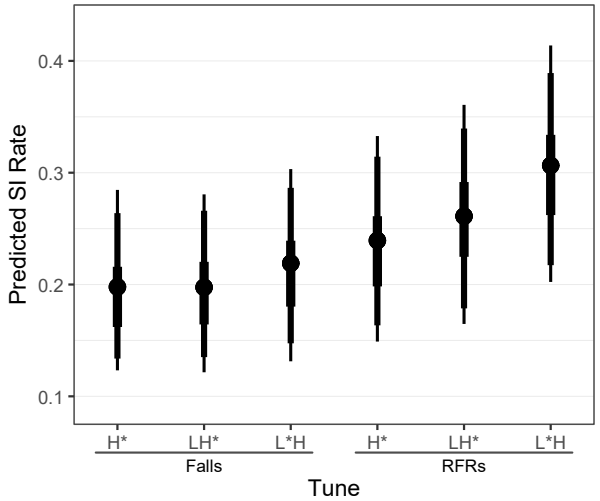


Figure 4.11: Exp. 4 (dual task) posterior-predicted mean SI rates with 50/89/95% highest density intervals.

Based on this figure, we can observe the same pattern within the RFR groups: SI rates show a numerical cline from H* to L+H* to L*+H with L-H% edge tones but not with L-L% edge tones. The statistical model results are shown in Table 4.8; as in Exp. 1, the RFR-shaped tunes have higher SI rates than the Falls ($\hat{\beta} = 0.36, CI = [0.08, 0.63]$).

Term	Estimate	Std.Error	95% CrI	PD
Intercept	-1.20	0.22	$[-1.64, -0.77]$	100.00
LH*−H*LL	0.00	0.21	$[-0.42, 0.42]$	50.42
L*H−H*LL	0.13	0.21	$[-0.27, 0.53]$	73.62
RFRs−Falls	0.36	0.14	$[0.08, 0.63]$	99.41
LH*−H*LH	0.12	0.20	$[-0.27, 0.50]$	72.17
L*H−H*LH	0.34	0.20	$[-0.06, 0.75]$	95.88

Table 4.8: Exp. 4 (dual task) statistical model results for the SI portion of the dual task results. Estimates are given on the log-odds scale.

4.4.5.3 Pooling SI Results with Exp. 1

Complementary to this analysis is the related question of the role of alternative salience in SI computation. Van Tiel et al. (2016, pp. 13–14), following a Neo-Gricean framework, hypothesize that in order for SI to arise, the hearer must reason about the informationally stronger alternative that the speaker could have said—thus the alternative must be available, or salient, to the hearer to reason about. Yet, "salience" and "availability" may be operationalized in different ways. Van Tiel et al. (2016) do not find evidence for an effect of availability based on lexical measures such as association strength, grammatical class, word frequency, or semantic relatedness. They additionally argue against the possibility that any possible effects of availability would be saturated by the fact that the higher alternative is present in the probe question (e.g., *Would you conclude that [...] not **cold**?*) as it is not necessarily the case that the hearer used this alternative when reasoning about the speaker's intents (because the question was not posed to the speaker). Schwarz et al. (2016) do not find evidence for an effect of availability using a subliminal priming paradigm, where the higher alternative (e.g., *cold*) is briefly shown for only 32-48ms prior to the probe question (i.e., below the conscious perceptible threshold of the participant). In contrast, Ronai & Xiang (2024) showed that

making the higher alternative available in the discourse context via the explicit QUD does increase SI rates. Similarly, Ronai & Göbel (2024) attribute the increased SI rates found with the use of RFR to increased salience or availability of the higher alternative following the presuppositional account of Göbel (2019).

Alternative "availability" in the dual task presented here can be seen as an extension of the subliminal priming task used by Schwarz et al. (2016), where rather than showing the higher alternative faster than the participant can consciously perceive it, the participant instead must perceive it and retrieve it from the lexicon in order to make the lexical decision prior to then being asked the SI-probing question. In this light, it is possible that making the higher alternative more available through the lexical decision task may increase SI rates.

As a follow-up analysis, the SI judgments from Exp. 1 (web-based auditory inference task) and Exp. 4 (in-person dual task) are pooled together to see whether the difference in single versus dual task set up—where the latter has increased the availability of the higher alternative—affected SI judgments. In addition to the terms from the previous model, the pooled model includes a fixed effect of Experiment[46] such that the intercept and effect of tune is averaged over the two experiments as well as an interaction with Tune. No interaction term showed a probability of direction over 81%, so only the fixed effects are reported here for brevity. The model results are shown in Table 4.9.

Term	**Estimate**	**Std.Error**	**95% CrI**	**PD**
Intercept	-1.09	0.14	$[-1.37, -0.82]$	100.00
LH*−H*LL	0.03	0.12	$[-0.20, 0.27]$	59.45
L*H−H*LL	0.00	0.12	$[-0.24, 0.24]$	50.38
RFRs−Falls	0.32	0.07	$[0.18, 0.46]$	100.00
LH*−H*LH	0.06	0.12	$[-0.17, 0.30]$	70.41
L*H−H*LH	0.27	0.11	$[0.05, 0.49]$	99.02
DualTask−SIOnly	0.09	0.14	$[-0.18, 0.35]$	74.00

Table 4.9: Statistical model results for the pooled SI results from Exp. 1 (SI Only) and Exp. 4 (Dual Task). Estimates are given on the log-odds scale.

[46]Experiment 1 (SI Only) is coded as -0.5 and Experiment 4 (Dual Task) is coded as $+0.5$).

The statistical model shows no credible effect of experiment ($\hat{\beta} = 0.09, CI = [-0.18, 0.37]$), suggesting that the additional availability from the lexical decision portion of the trial did not serve to increase SI computation. Notably, the improved statistical power from pooling the data from the two experiments together shows stronger evidence of a difference between L*+HL-H% and H*L-H% ($\hat{\beta} = 0.25, CI = [0.03, 0.47]$). As seen in the previous analyses of the two experiments, there remains a main effect of RFR versus Falling contour shape such that RFR-shaped tunes as a group show greater SI rates than the Falls ($\hat{\beta} = 0.31, CI = [0.17, 0.45]$).

4.4.5.4 Lexical Decision Results

Overall, RTs tended to be longer in the dual task compared to the stand-alone lexical decision task (Exp. 2) and so the same exclusion criterion used in the online experiment is used here, leading to a data loss of 0.98%. The statistical analysis, using the same model structure as in the previous experiments, show that results are overall similar to what was seen in the online experiment. Namely, there are robust effects of semantic priming ($\hat{\beta} = 0.063, CI = [0.038, 0.088], PD = 100.00$) and contrastive versus non-contrastive alternatives ($\hat{\beta} = 0.024, CI = [0.000, 0.048], PD = 97.65$), but the by-tune results are seemingly not robust to the dual task setup—the full model table and by-tune results figure appear in to Table C7 and Figure C18 in the Appendix. One conjecture is that because the task demanded participants to switch between a speeded lexical decision and an unspeeded inference judgment, there was a "switch cost" in the task which made the RTs overall longer and more variable.[47]

Recall that one limitation of the previous lexical decision experiments, and indeed previous work on the processing of scalar alternatives using lexical decision more broadly, is that there was no way to know whether participants were computing the SI-enriched interpretation on any particular trial when they made their lexical decision response. The dual task thus provides us the distinct opportunity to condition RTs based on the participant's response on the inference task portion of the trial. Participant RTs were modeled using a second statistical model, now including

[47] See Figure C13 and Figure C20 in the appendices for a comparison of RT distributions.

a predictor of whether the participant responded Yes or No to the comprehension question on each trial. The posterior predicted differences between each condition are shown in Figure 4.12 (see Figure C19 in Appendix C.7 for the by-condition predicted RTs).

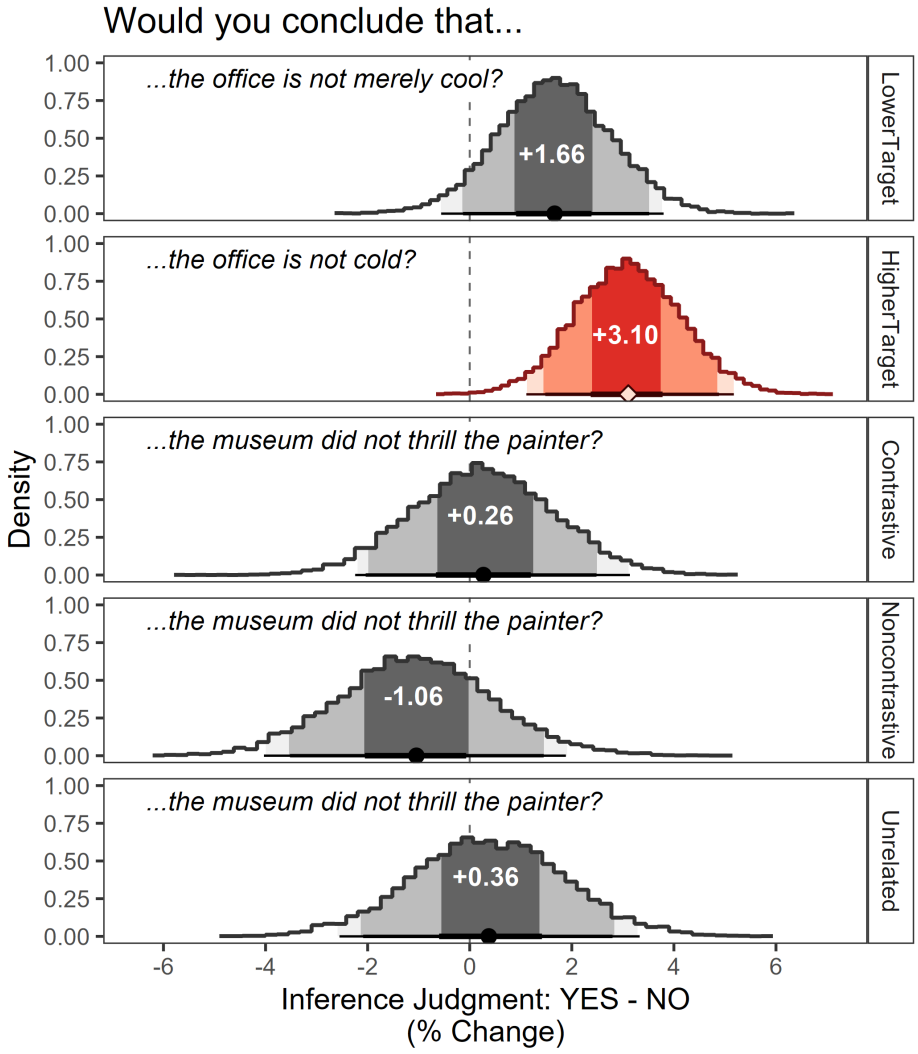


Figure 4.12: Exp. 4 (dual task) posterior distribution of the difference between when participants respond Yes versus No in each target condition. Distribution reflects the percent change in RT relative to a No response; negative values indicate RTs are faster when the inference is drawn while positive values indicate that RTs are slower when the inference is drawn. In the HIGHERTARGET condition, a response of Yes indicates that the participant computed SI. See (17 (p. 141)) for examples of the items.

The model finds a credible difference[48] for only the HIGHERTARGET condition. In other words,

[48]Due to the size of the statistical model, for brevity here only the manually-computed posterior estimates for the contrast between Yes and No responses in each condition are reported rather than any particular coefficient. This is

when participants respond Yes in the condition that probes SI, their RT for the lexical decision portion that preceded the question was higher ($\%\Delta = +3.10, CrI = [+1.11, +5.17], PD = 99.88$). This effect does not show an interaction with any specific tune, and none of the differences within the other conditions are credibly different from 0. Note that this runs counter to typical expectations about the direction of effect expected for SI processing (e.g., assumed by Lacina et al., 2024; Ronai & Xiang, 2023 i.a.), where the prediction is that greater likelihood of SI is associated with facilitation. Although, this result is generally in line with the finding in Exp. 2 (long SOA lexical decision) where it was found that the RFR associated with the greatest likelihood of SI-enrichment was also associated with less facilitation. This result is also in line with a recent (and similarly unexpected) finding from Lacina & Gotzner (2024), where lexical scales with higher SI rates showed slower RTs in a text-based lexical decision task—i.e., there was an negative relationship between propensity for SI enrichment and priming. These results will be discussed further in the general discussion.

4.5 General Discussion

This work has presented five experiments to investigate whether different RFR-shaped tunes behave differently or similarly in offline interpretation (via the inference task) and in online processing (via the cross-modal lexical decision task). The potential differences between tunes in the context of SI as it relates to adjectival scales were investigated by operationalizing the notion of "higher alternative" as the higher scalemate of a lexical scale. Three research questions regarding how alternatives are processed and how RFR interacts with different types of alternatives were investigated.

The first question was whether RFR-shaped tunes behave similarly or differently from one another. In offline interpretation (discussed further in §4.5.1), it was found that the three RFR-shaped tunes as a group showed higher SI rates compared to falls. Moreover, there was a cline between the three RFR-shaped tunes such that H*L-H% had lower SI rates compared to L*+HL-

the same approach used to compute and plot the by-tune results earlier in this work, which have already been shown to correspond to the model parameters when the contrasts are set up appropriately.

H%, with L+H*L-H% lying somewhere in between. In online processing (discussed further in §4.5.2), the RFR-shaped tunes behaved differently from one another, where H*L-H% showed a greater degree of facilitation of the higher alternative while L*+HL-H% instead showed less facilitation.

The second question was whether scalar alternatives behave differently from focus alternatives in processing (i.e., setting aside further modulation from individual tunes). From the lexical decision results, it was shown that there was no credible difference between the CONTRASTIVE versus the LOWERTARGET and HIGHERTARGET conditions. There was additionally no credible difference between the LOWERTARGET and HIGHERTARGET conditions overall, in contrast to the prediction from De Carvalho et al. (2016) that there would be an asymmetry. These results are discussed further in §4.5.2.

The third question was whether there was an asymmetry in the processing correlate of RFR such that it specifically targets the higher, and not the lower, alternative. Such an asymmetry was found in the late SOA lexical decision task (Exp. 2). Specifically, it was found that no intonational tune modulated the activation status of the **lower** alternative, but the RFR-shaped tunes did modulate the activation status of the higher alternative. As previously mentioned, L*+HL-H% showed a greater degree of facilitation of the higher alternative while L*+HL-H% instead showed less facilitation. In contrast, the results of the early SOA experiment (Exp. 3) showed that no intonational tune modulated the activation status of the **higher** alternative, but L*+HL-H% showed a greater degree of facilitation while H*L-L% showed less facilitation. A more detailed discussion of the difference in effect direction for Exp. 2 and the puzzling pattern of results for Exp. 3 is discussed in §4.5.4.1.

4.5.1 Discussion of Inference Task Results

This work has presented a series of inference tasks probing the likelihood of SI-enriched interpretations with indirect question-answer pairs. In a text-only norming task, the rate of SI computation varied across scales, replicating prior work on scalar diversity (i.a. Gotzner et al., 2018; Ronai &

Xiang, 2024; Sun et al., 2018; Van Tiel et al., 2016). While prior work has also examined scalar diversity in dialogue contexts (Ronai & Göbel, 2024; Ronai & Xiang, 2024), the current experiments differed from prior studies in that the indirect question-answer dialogues used here did not mention the higher alternative explicitly in the dialogue. In an inference task using audio which additionally manipulated the intonational tune used with the answer (Exp. 1), RFR-shaped tunes showed greater SI rates compared to Falls. Within the RFR group, there were numerical differences between the three pitch accents such that L*+H showed the greatest SI rates followed by L+H* then H*. In a dual task paradigm containing an inference task following a lexical decision of the higher alternative (Exp. 4), both of the results of Exp. 1 were replicated. When pooling the data together in a single statistical model, there was stronger evidence for a cline across the rising accents such that L*+HL-H% had higher SI rates than H*L-H%, with L+H*L-H% lying between the two. Moreover, there was no difference in SI rates between the two experiments. In summary, the main finding from the inference task results is that RFR-shaped tunes overall behave alike in offline interpretation in the domain of SI computation, with small numerical differences between the pitch accents. The remainder of this section will relate these results to ongoing discussions in the SI literature.

4.5.1.1 Regarding Alternative Salience

Prior work on SI has ascribed varying levels of importance to the role of alternative salience or availability in SI computation. Some researchers argue that sufficient salience of the higher alternative will necessarily lead to SI computation, at least in the context of structural priming paradigms (Rees & Bott, 2018, see similar claims by Bott & Chemla, 2016).[49] Others find evidence either that this view is too strong (Waldon & Degen, 2020) or incorrect (Marty et al., 2024), the latter arguing instead that salience does not play an independent role distinct from contextual relevance of the higher alternative and adaptation effects over the course of the experiment (see also discussion in

[49]For instance, Bott & Chemla (2016) show that when first presented with a card where all the shapes are stars, participants are more likely to calculate the "some but not all" SI on a later card; here, the notion of "salience" regarding *all* in their paradigm is different from overtly presenting participants with the lexical item *all* in the visual presentation of the experiment (as was done in the dual task presented here).

Ronai & Xiang, 2024, p. 22). Similarly, Schwarz et al. (2016) report that subliminal priming of the higher alternative (i.e., briefly displaying the higher alternative for 32-48ms) does not increase the rate of SI computation.

The dual-task (Exp. 4) results can be seen as an extension of Schwarz et al. (2016), where the "priming" of the higher alternative prior to presenting the SI-probing comprehension question for the inference task is not subliminal but overt: The participant is required to retrieve the higher alternative from the lexicon in order to make the lexical decision. However, when compared to the results of Exp. 1, which had no lexical decision portion prior to the inference judgment, there was no effect of experiment, speaking against an independent role of mere salience of the alternative. This result is also in line with the finding from Van Tiel et al. (2016) that lexical-based measures of availability did not make a strong impact on SI rates. On the surface, these results seem to be against the salience-based account described by Göbel (2019), who provides a presuppositional account such that the use of RFR increases the salience of the higher alternative via presupposing its existence.

There are two ways to view these seemingly contradictory results where salience either does or does not show an effect on the likelihood of SI calculation. One view, following Rees & Bott (2018), is that some experimental manipulations do not affect the salience of the higher alternative to a high enough degree to impact the overall rate of SI computation. That is, if SI necessarily arises if the higher alternative is sufficiently salient, then perhaps subliminal priming and lexical retrieval (via lexical decision) does not make the higher alternative fully salient for the inference to be robustly computed.

An alternative view may be that it is not mere cognitive salience that matters for SI computation, but that it is instead the contextual relevance of the higher alternative that matters. Taking the view of Marty et al. (2024, p. 2), the higher alternative is more contextually relevant "when an utterance of an expression is able to address a question, whether explicitly raised or implicit in the context, the information encoded in the expression, as well as the expression itself, is described

as being relevant to the [QUD]."[50] Prior work has shown that the likelihood of SI computation increases substantially when the higher alternative is included in the QUD (Ronai & Xiang, 2024), where the higher alternative is not only merely salient (e.g., because the word *cold* is presented) but it is also contextually more relevant by virtue of being explicitly queried in a way that is not present with mere subliminal priming or lexical retrieval. So, while Ronai & Göbel (2024) propose that RFR increases the salience of the higher alternative, it may perhaps be more apt to describe this as establishing the higher alternative as contextually relevant for the interpretation of the utterance in much the same way as focus-marking conveys that alternatives to the focused constituent are relevant to interpretation.

4.5.2 Discussion of Priming Results

This work presented four cross-modal lexical decision tasks probing the processing profile of different tunes as they relate to higher and lower alternatives. The primary prediction was that if RFR makes SI more likely by virtue of its relation to higher alternatives, then there should be increased facilitation (=faster RT) for the higher alternative when RFR is used with the lower alternative. For example, *cold* will be more activated upon hearing *cool* uttered with RFR than if it were uttered with another tune. The real question here is whether all RFR-shaped tunes would behave this way or whether only one does.

The main finding is that, unlike the inference task results, the RFR-shaped tunes behave **differently** from one another in online processing at different time points. Early in processing (Exp. 3, using a 0ms SOA), there is increased facilitation of the lower alternative when L*+HL-H% is used and no differences in processing for the higher alternative with any tune. Later in processing (Exp. 2, using a 750ms SOA), the same tune (L*+HL-H%) leads to less facilitation of the higher alternative while H*L-H% leads to more facilitation. Unlike the early SOA processing profile, there is no difference in processing for the lower alternative with any tune. When the experimental

[50] Marty et al. (2024, p. 2) provide an example where the SI-enriched interpretation of *Some of the symbols are circles* is more robust with the QUD *What symbols are on that card over there?* than with *Are there any circles on that card over there?*.

environment is relatively uncontrolled in an online setting (Exp. 2b) or part of a dual-task setup (Exp. 4), the effects of intonational tune are not robust—this point will be discussed further in the limitations section.

The results of the lexical decision tasks additionally showed that scalemates offer no processing advantage early or late in processing compared to contrastive alternatives when controlling for word length and frequency. One may object that the contrastive relation between *cool* and *cold* or *sculptor* and *painter* is only invoked when the accented element is contrastively accented—in other words, the distinction is only relevant when L+H* is used. As mentioned in the interim discussion of Exp. 2, a targeted comparison of just the L+H*L-L% between the scalar conditions is not feasible because tune did not vary by item in the HF16-adapted item set (i.e., only 10 items appeared in the L+H*L-L% condition, c.f. all 64 items in the critical item set). This caveat aside, there was nonetheless a distinction between the CONTRASTIVE and NON-CONTRASTIVE conditions overall when averaging over all intonation conditions, suggesting that there is a processing advantage for items with the potential to be contrastive alternatives that is not present for non-contrastive associates. However, there was no difference between the CONTRASTIVE condition and the two scalar conditions (the LOWERTARGET and HIGHERTARGET conditions). This result suggests that while not all (potentially) contrastive associates are scalemates, those associates that happen to be scalemates do not show a unique overall processing advantage. In other words, the facilitation of the scalar alternatives appears analogous to the focus alternatives. This finding is in stark contrast to the asymmetry reported by De Carvalho et al. (2016) where weaker scalar terms prime their stronger scalemates more than the reverse. More broadly, this result speaks against accounts where SI is automatic such that the stronger term is necessarily required in the interpretation of the weaker term.

This work also presented a dual task combining the inference task and the lexical decision task to explicitly relate patterns in priming to participants' interpretations. Experiment 4 sought to address a limitation in prior priming studies involving scalar alternatives (De Carvalho et al., 2016; Lacina & Gotzner, 2024; Ronai & Xiang, 2023), where patterns in priming could not be

directly related to whether SI was computed or not; although, it should be noted that this is also a limitation of lexical decision tasks with focus alternatives, as it is not directly probed whether participants excluded the specific alternative that was probed. Though the taxing nature of the dual task setup (from switching between the two tasks) inadvertently made identifying by-tune effects difficult, as evidenced by an increase in RT variability, RTs were nonetheless higher in the HIGHERTARGET condition when participants responded "Yes" on the subsequent SI-probing question. As previously mentioned, this result was unexpected but not without precedent, as Lacina et al. (2024) also found a similar negative correlation in a between-experiment comparison of SI rates and degree of priming.

The relationship between higher SI rates yet a seemingly lower degree of priming found in the dual task (as seen in Fig. 4.12) warrants future research explicitly relating the activation status of (focus or scalar) alternatives to participants' actual interpretation, it can nonetheless be viewed through the lens of previous work reporting that SI is costly (Bott & Noveck, 2004; De Neys & Schaeken, 2007; Huang & Snedeker, 2009, among many others, though importantly c.f. Degen & Tanenhaus, 2015; Grodner et al., 2010 i.a.). It is plausible that such a processing cost, which has previously been shown through a variety of measures, slows RT in lexical decision tasks. As a result, any potential facilitation of alternative activation may be counteracted by the cost of SI. It is crucial to note, however, that such a processing cost in the context of SI would need to be reconciled with not only the overall finding presented here that focus and scalar alternatives are not different from one another, but also with the general finding that effects of focus are generally facilitatory (Gotzner & Spalek, 2019). The pattern of results found here when conditioning the RT results on the participants' reported interpretations is in need of further validation, potentially through a different paradigm or a restructuring of the trial structure for the dual task to avoid "switch costs" between the two task components.

As mentioned in the interim discussion following Exp. 2 (long SOA lexical decision), the priming results for the RFR-shaped tunes are somewhat counterintuitive given the SI results previously discussed. L*+HL-H% yields the highest SI rates, yet yields less facilitation of the higher al-

ternative later in processing, while H*L-H% yields comparably lower SI rates, yet yields more facilitation—these relationships are the opposite of what was predicted. A cursory glance at these patterns may suggest that the negation of the higher alternative is characterized not by activation but suppression, as it is no longer being considered. However, such an inhibitory process would run counter to the broader literature showing (via a variety of experimental methods) that focus alternatives in fact have strengthened representations, hence yielding facilitatory effects (i.a. Fraundorf et al., 2010; Gotzner & Spalek, 2019; Gotzner et al., 2016; Spalek et al., 2014). Even more unexpected is that the L*+HL-H% tune yields more facilitation of the lower alternative (while H*L-L% yields less) early in processing whereas the higher alternative shows no sensitivity to the intonational tune.

4.5.3 Relating the Results to Formal Accounts

How do the results from the experiments reported here relate to existing accounts of RFR? Generally, the SI results replicate prior work from de Marneffe & Tonhauser (2019) and Ronai & Göbel (2024) and are in line with the account from Göbel (2019), though with the previously mentioned caveat of either reframing salience as contextual relevance or by differentiating the degree of salience across different types of experiments. As mentioned by de Marneffe & Tonhauser (2019), the account of Constant (2012) would likely need to be weakened somewhat to account for how SI rates increased with the RFR-shaped tunes, but were nonetheless relatively low.

With regard to the UNCERTAINTY account of Ward & Hirschberg (1985), the SI results presented here speak against a type of uncertainty that requires the truth of the higher alternative to remain unknown. Although, because the dialogue contexts used in this work did not contain the lexically-specified higher alternative, it is possible that other types of uncertainty described by Ward & Hirschberg (1985) may remain as plausible descriptive accounts of RFR.[51] For example, a speaker may convey uncertainty about whether an ad-hoc scale between <*cool*, *window left open*>

[51]Moreover, it is left unexplored here whether 'uncertainty' should be narrowly associated with the primary implicature ($\neg K\phi$) described by Sauerland (2004) or the ignorance inference ($\neg K\phi \wedge \neg K\neg\phi$) described by Buccola & Goodhue (2023).

(which are not related by asymmetric entailment) is relevant. The results from the online processing experiments in the present study are nonetheless compatible with the incredulous interpretation described by Hirschberg & Ward (1992).

The results of the present study do not support accounts that appeal to broader QUDs like that of Büring (2003) which predict lower SI rates compared to falls. The secondary QUD account of Westera (2019) and the incomplete answer account of Wagner et al. (2013) would be compatible with either an increase or decrease in SI rates, and so they are supported insofar as there is a difference between falling and RFR-shaped tunes (i.e., there is a difference between the H% and L% boundary tones).

4.5.4 Relating the Results to Phonological Theory

Up to now, this work has placed rather strict boundaries between the tunes under investigation on the basis of the AM model for MAE—namely, a categorical distinction between rising pitch accents. But the results of this work do not show consistent patterns within a pitch accent when used with either edge tone. Rather, the observed effects related to modulation of SI rates and participant RTs are only ever seen when considering the entire tune. These results thus speak to a need to consider variation at the level of the holistic tune rather than merely the choice of pitch accent divorced from its edge tone context. In terms of the experimental materials, note that H*, L+H*, and L*+H all adhere to prior phonological and phonetic descriptions but as a result co-vary in terms of pitch range. Recasting the materials in this light, rather than a clean divide between three distinct RFR-shaped tunes, the materials can be described as a three-step continuum between a low-scaled RFR (H*) and a high-scaled RFR (L*+H)— all part of a larger class of RFR-shaped tunes.

What motivation beyond these results is there to loosen the divide between rising accents? As described in the introduction, the categorical status of H* versus L+H*, which is commonly taken as a given in experimental pragmatics (often referred to as “neutral” versus “focus” prosody, e.g., Husband & Ferreira, 2016; Rett & Sturman, 2020), is much more controversial in the intonational

phonology literature (Calhoun, 2006; Gussenhoven, 2016; Ladd, 2008, 2022; Ladd & Morton, 1997) and is among the most difficult distinctions made in a ToBI annotation (Pitrelli et al., 1994; Silverman et al., 1992; Syrdal & McGory, 2000). In production, the choice of rising pitch accent is probabilistically associated with information structure (Im et al., 2023; Roessig, 2021) rather than deterministically associated with the presence or absence of narrow focus. In perception, H* and L+H* overlap in their capacity to convey contrastive focus (Roettger et al., 2019; Watson et al., 2008). Moreover, the distinction between rising accents is often limited in production and discrimination between them is poor in most edge-tone contexts (Steffman et al., 2024). In recent modeling work, the inventory of pitch accents has been alternatively described as arising from a singular dynamical system, where continuous variation in one parameter leads to the observed differences in trajectories for the three rising accents (Iskarous et al., 2024)—that is, the three rising accents can be viewed as a continuum, spanning trajectories between H* and L*+H.[52] The results from the current work suggest that the seemingly categorical distinctions may be more aptly described in terms of meaningful gradience (Ladd, 2008, pp. 151–156 and Ladd, 2022, pp. 252–253) within a single category, rather than cleanly dividing three separate RFR-shaped tune categories.

4.5.4.1 Applying Pitch Range to the Present Results

The seemingly paradoxical relationship between raised SI rates and lowered degree of facilitation among the RFR-shaped tunes may perhaps be explained in terms of variation in pitch range within a broad RFR-shaped class. In the context of the results of this work, the RFR with the largest pitch range, L*+HL-H%, may be inviting additional inferences while H*L-H%, which has a much smaller range, does not. These additional inferences, which likely arise later in processing, in turn contribute to the slower processing profile of L*+HL-H%. L+H*, whose pitch range lies somewhere in the middle, subsequently falls in-between the faster H*L-H% and the slower L*+HL-H%. While the content and cardinality of what these additional inferences may be are left unexplored

[52]Technically speaking, under this model, increasing values of the stiffness parameter k continuously varies the rise trajectory from H* to L+H* to L*+H, but past a certain value the trajectory changes to yield falling accent trajectories (L*, H+!H*) rather than rising trajectories.

here, one likely competing inference would be that of INCREDULITY. In seminal work on incredulity and RFR, Hirschberg & Ward (1992) originally argued that RFR is compatible with two interpretations, either UNCERTAINTY (i.e., a lack of speaker commitment) or INCREDULITY, with the latter being more likely when RFR is phonetically implemented with increased pitch range.[53] If an analysis rooted in gradient distinctions related to pitch range is on the right track, then the results here also provide support for observations from Westera (2019, p. 326) and Constant (2014, p. 27), who claim that differences between RFR-shaped tunes likely arise from paralinguistic meaning conveyed via gradient scaling and alignment differences. Under this view, the expected facilitation from RFR (which is found with H*L-H%) may be masked as pitch range is made more extreme (as seen with L*+HL-H%).

This account assumes that the effects of RFR and pitch range come later in processing, and so it is a puzzle why there is increased facilitation of the **lower** alternative early on in processing with L*+HL-H%. Under an incredulous view, this may perhaps come as a result of a need for a lower standard of comparison with which to compare the invoked scalar value, leading to increased relevance of lower values like *cool* when expressing incredulity about *cold*. What is perhaps more interesting about this result though is that variation in processing time related to tune is in the opposite condition than in the later SOA experiment. Early in processing, only the lower scalar values, which were entailed by what was uttered (by virtue of operationalizing the scales here to entailment-based scales), showed variation in response to the tunes—the higher values which are not entailed did not show any pattern across the tunes. Later in processing, the lower scalar values show no pattern while the higher alternatives show the expected pattern in relation to RFR (*modulo* the previous discussion of pitch range). Existing models of RFR focus almost entirely on secondary/broader/higher alternatives, so it is not clear why there would be an early sensitivity to

[53] Similar conjectures can be found in other works on RFR; for example, Constant (2014, 27, fn 19) offers the suggestion that differences in RFR-shaped tunes "can be understood in terms of a gradient paralinguistic effect where [L*+H] is perceived as more 'emphatic.'" Westera (2019, p. 326) is explicit in treating different RFR-shaped tunes as the same, with the suggestion that the delayed peak in L*+H may "indicate extra significance" and that the distinction between UNCERTAINTY and INCREDULITY is related to paralinguistic variation. To these authors, it may come as a welcome empirical validation that such variation in pitch range can account for the initially counterintuitive findings across the experiments presented in the present work.

intonation for the lower/entailed scale value; this puzzle is left for future research.

4.5.5 Limitations and Future Work

While there are ways in which the predictions of different formal accounts might be speculatively cashed out in terms of the presence/magnitude of priming, these accounts do not lay out processing predictions explicitly. As a result, the priming results primarily speak to the assumption that RFR relates (asymmetrically) to higher rather than lower alternatives. It was predicted that if there were to be an effect of RFR on a prime for lexical decision, it would be one of facilitation analogous to findings on the processing of focus alternatives. An avenue towards adjudicating prior formal accounts on the basis of lexical activation is left to future work.

While the magnitude of the priming effects reported in the lexical decision tasks (Exp. 2-4) are small, they are in fact in line with prior work from Husband & Ferreira (2016, p. 226). In their results, the difference between intonation conditions were only percent changes of about 0.08% to 2.3%. In this work, the most robust effect was that of semantic priming (comparing semantically unrelated to related words), which showed an effect size of 5.46% to 9.05% depending on the experiment. So, the effects reported here of $\approx$2% are well within the reasonable effect size that could be expected. It should additionally be emphasized that the goal of this work was not to precisely measure the magnitude of facilitation; for instance, the goal was never to discover that RFR speeds up lexical decisions by precisely 17ms. Rather, inference for the by-tune effects was operationalized primarily on probability of direction: Is there evidence in favor of generally faster or slower responses, as measured via percent change, beyond what is already accounted for by semantic association. It was also shown that these by-tune effects are not robust to (1) experiment-extrinsic sources of noise such as uncontrolled environments (as shown by the online lexical decision results) and (2) "task switch" effects when going from an unspeeded judgment to a speeded judgment (as shown by the dual task results). The takeaway from this is that because the effect of intonation is subtle in this paradigm, future work investigating these subtle by-tune

effects[54] should be done in a very carefully controlled laboratory environment.

Regarding the time course of processing, the lexical decision results presented here only probe an early (0ms) and late (750ms) timepoint. Future work would likely benefit from exploring other methods such as pupillometry or eye tracking to investigate the entire timecourse of processing. The former may be particularly promising, as it would not require training participants to associate specific pictures with specific scale values for a large number of scales as would be needed for a study of multiple different adjectival lexical scales in the visual world eye tracking paradigm.

Regarding scalar diversity, it was not an aim of this work to uncover factors that explain variation in the phenomenon. Rather, the existence of such variation was taken as a given, with the knowledge that intonation has already been shown to modulate SI calculation rates. This work replicates previous results in this domain: The use of RFR-shaped tunes increase the rate of SI computation compared to falls (Ronai & Göbel, 2024). However, it is nonetheless informative that the scalar diversity finding replicates (across three experiments) with the indirect question-answer pairs, as previous QUD manipulations have involved either no QUD (i.e., just the SI-probing comprehension question) or a QUD with the stronger alternative.

Whereas this work uses indirect question-answer pairs to avoid saturation effects in the RT measurements of the lexical decision task, future work may additionally consider whether the magnitude of the effects found here interacts with the QUD of the dialogue. For example, for testing SI without the addition of a parallel lexical decision task using the same stimuli, the QUD *Did someone leave a window open in the office overnight?* could be straightforwardly changed to something like *Does the office feel <u>cold</u> today?*, which explicitly makes *cold* at-issue. The response *The office feels <u>cool</u>* would then be expected to have overall higher rates of SI compared to the experiments presented here (see also Ronai & Göbel, 2024) and the numerical cline observed in the RFR-shaped tunes may be more apparent in magnitude.

Related to the constraints in the materials, it should be noted that this work exclusively inves-

[54]Note that this is not intended to generalize to all online priming tasks, as some effects (such as mere semantic priming) were evident even in the online experiment. In other words, there are some effects in the lexical decision paradigm that are robust to imperfect experimental conditions and others that are not—effects of intonation are likely in the latter category.

tigated scales defined via asymmetric entailment (Horn, 1972) and did not seek to address scales that can be defined using relations other than entailment. For instance, Göbel (2019) describes RFR in relation to evaluative scales as in (19), where RFR is used to relate a positive evaluation to an already salient negative evaluation.

(19) A: Dexter is such a horrible person! (Göbel's 15a)
B: He gives to *charity*...

Ward & Hirschberg (1985) also offer numerous spontaneous examples of RFR that are felicitous yet are not restricted to lexical scales as in (20), where the scale is defined with an 'is a part of' ordering relation.

(20) A: Did you read the first chapter? (Ward and Hirschberg's 25)
B: I read the first *half* of it...

Although reading the first chapter entails reading the first half of it, hence establishing an entailment relation that is not lexically specified (c.f. <*some*, *all*>), such an entailment relation is not necessary. For example, RFR is also used felicitously in (21) even though going to Tokyo does not entail going to Paris.[55] An ordering relation defining a scale between *Tokyo* and *Paris* (such as 'is more desirable to go to') can still be easily considered for (21). Yet, RFR can nonetheless be used felicitously in examples like (22), where it is not straightforward to determine what the relevant scale should be between *Ellen's* and *Midtown*.

(21) A: Did John go to Tokyo?
B: He went to *Paris*...

(22) A: Have you ever had dinner at Ellen's? (Ward and Hirschberg's 34)
B: We've had lunch at *Midtown*...

If one were to use lexical decision to probe examples like (19–22), where the scales are not lexically specified, the target word for the task would need to be a word that has already appeared

[55]Thank you to Gregory Ward (p.c.) for the example.

in the discourse.[56] Again, this situation where the target is explicitly mentioned in the discourse is precisely what the materials presented in §2.2 sought to avoid. So, these kinds of examples would not easily fit into the current lexical decision tasks. Future work may opt to relax this constraint in the materials or adjust the scope of the experiment to investigate a wider variety of contexts that support the felicitous use of RFR without the express need of a lexically-specified entailment relation. For example, one could conceive of an experiment comparing (unordered) focus alternative sets to (ordered) scalar alternative sets.[57] Moreover, although the contribution of the RFR-shaped tunes investigated here was argued to be linked to a broad RFR class with gradient phonetic variation, it may yet be possible that the RFR-shaped tunes are categorically different from one another in contexts that are not constrained by lexically-specified entailment-based scales.

Regarding potential non-F0 cues to the intonational tunes here, the goal with creating the materials was to allow for such covarying cues (e.g., duration, see Sandberg, 2024 and Arvaniti et al., 2024) to exist in the signal while controlling for much of the idiosyncratic variation in F0. This approach is in contrast to selecting a single source recording and resynthesizing six different contours from it. Anecdotally, there are durational differences between the six tunes; for example, contours with L*+H tended to have longer duration than contours with H*. While this may present a confound where longer tokens may afford greater activation than shorter tokens, this hypothesis would predict the highest levels of facilitation for L*+H for both the HIGHERTARGET and LOWERTARGET conditions in each experiment—such a pattern was not found. The only instance where L*+HL-H% yielded the fastest RTs was in the short SOA experiment for the LowerTarget condition; in all other conditions and experiments L*+H tunes are either no different from the condition mean or are slower (as in the main result for the long SOA task, or numerically in the dual task for both conditions). A more targeted investigation of the systematicity of non-F0 cues across the six

[56]The prime-target pairs would need to be *charity-horrible* (for 19), *half-chapter* (for 20), *Paris-Tokyo* (for 21), and *Midtown-Ellen's* (for 22).

[57]For instance, one could probe the activation of *chapter* in (20) in a "neutral" condition (with H*L-L%), a "focus" condition (with L+H*L-L%), and an "evoked scale" condition (with RFR). Probe recognition, rather than lexical decision, may potentially be less sensitive to saturation effects from overt mention of the target (see i.a. Gotzner & Spalek, 2019; Muxica & Harris, 2025 for examples of such tasks).

tunes is left to future large-scale production studies, as the peak location/height analysis presented here is limited to recordings from only a single speaker.

This work looked at falls (L-L%) and RFR-shaped tunes (L-H%), and so it may seem tempting to pin a common difference on the distinction in edge-tone configuration (or more narrowly, the final boundary tone). However, the one missing piece in doing so is the L* pitch accent, which was not included here but of course can combine freely with both edge-tone configurations. Yet, L*L-H% is, by definition, not an RFR-shaped tune because L* is not a rising accent—it is a falling accent. The only times L*L-H% is mentioned in the RFR literature is when comparing accounts to the CONTRADICTION CONTOUR (Liberman & Sag, 1974), which is sometimes described as being different from RFR given that the choice of L* versus L*+H does not appear to matter (Constant, 2012, c.f. Westera, 2019). Whether the common core for the three RFR-shaped tunes can be pinned down to the L-H% edge-tone configuration, thus extending to L*L-H%, remains an open question unexplored here. Those interested in a compositional theory of intonational meaning may find it worth exploring further while those against may be content with a broad distinction between Falls and RFR-shaped tunes.

4.6 Conclusions

Prior work on RFR varies as to (1) how RFR is defined on phonological grounds (2) what the pragmatic contribution of RFR is and (3) the extent to which diferent RFR-shaped tunes share a common core in terms of their meaning contribution. This work investigated different RFR-shaped tunes that differ in the pitch accent specification in the context of SI, which offers an operationalization of higher versus lower alternatives in the form of lexical scalemates such as $<$*cool*, *cold*$>$. The goal was to assess whether all, or only one, RFR-shaped tune is sensitive to the distinction between higher versus lower alternative. This was examined from two angles: offline interpretation using an inference task probing SI calculation and a cross-modal lexical decision task using carefully controlled auditory materials and equipment.

The results showed a counterintuitive relationship between SI rates and degree of facilitation:

Progressively higher SI rates in RFR-shaped tunes led to progressively less facilitation of the higher alternative in the lexical decision task later in processing. It was proposed that this relationship can be explained by considering the RFR-shaped tunes as part of a broad RFR class with meaningful variation in terms of pitch range such that a higher-scaled RFR may be more likely to invite other competing inferences such as an incredulous interpretation (Hirschberg & Ward, 1992). Methodologically speaking, the priming results are subtle and not robust to uncontrolled environmental factors, and so the use of web-based cross-modal lexical decision tasks when investigating intonation specifically is discouraged. Lexical-level differences such as general semantic relatedness are robust to such uncontrolled factors, though. Similarly, a novel dual-task setup with both lexical decision and inference tasks was used. While the RT data appeared to be compromised, the dual task nonetheless replicated the cline of SI rates for the RFR-shaped tunes found in stand-alone inference tasks.

The results of this work suggest that there is just one RFR shape but that within-category variation in pitch range may serve to convey other communicative functions that likely share a common core. However, whether this common core is specifically something like uncertainty, CT-marking, focus, or incompleteness would require a targeted investigation in isolating these different pragmatic phenomena and finding a way to accurately measure differences in participant interpretations given that participants are quite accommodating of intonation even between broad classes like Falls and RFRs. That said, the results presented here appear to be less compatible with uncertainty accounts and more compatible with other accounts. However, future work specifically probing speaker certainty may be able to further distinguish between pragmatic accounts of RFR. More importantly, these results suggest that there are not clean distinctions between the RFR-shaped tunes on the basis of the pitch accent, lending credence to claims in phonological theory positing meaningful gradience within a smaller set of contrastive intonational categories (e.g., Cole et al., 2023; Gussenhoven, 2004, 2016; Iskarous et al., 2024; Ladd, 2008, 2022).

Chapter 5

GENERAL CONCLUSIONS

5.1 Summary of Findings

The present thesis investigated two phenomena in MAE intonation: rising declaratives and RFR. Work on these phenomena are largely disjoint from one another and investigate different aspects of each tune, but they share an open question regarding whether there are meaningful distinctions to be drawn at the phonological level or whether there is meaningful gradience within broader tune categories. Part 1 (Chapters 2 and 3) investigated whether the distinction between INQUISITIVE rising declaratives and assertive rising declaratives was best characterized by a phonological distinction based on the starting point of the rise (i.e., the pitch accent, as posited by Jeong, 2018) or whether variation along this meaning distinction was better explained by other phonetic measures. Part 2 (Chapter 4) investigated whether participants responded similarly or differently to different RFR-shaped tunes in order to understand whether the connection to higher alternatives attributed to RFR is linked to a single RFR or was characteristic of a broader class of RFR-shaped tunes.

The findings from Part 1 showed that the ending F0 target was the primary cue associated with variation in INQUISITIVE/ASSERTIVE interpretation. Variation in the choice and scaling of the pitch accent (i.e., accentual pitch) played an inconsistent role and in the one instance where higher accentual pitch increased the likelihood of a TELLING response, the magnitude of the effect was substantially lower than that of ending pitch. These findings speak against a contrast between L*H-H% and H*H-H% on the basis of the INQUISITIVE/ASSERTIVE distinction, though it remains possible that these two tunes nonetheless contrast along some other dimension of meaning (see Gussenhoven & Rietveld, 2000 for a comparable investigation in Dutch with regard to speaker surprise). More broadly, the results are also compatible with a view that variation in the interpretation of rises is perhaps better characterized in terms of meaningful phonetic gradience rather than strictly categorical tune-level distinctions. Under this view, the results of Part 1 show what appears to be a prototypical INQUISITIVE rise: one that starts low and early and ends high;

deviation away from this shape–especially in ending F0–leads to a lower probabilistic likelihood of an INQUISITIVE interpretation. The followup analyses presented in Chapter 3 lend additional credence that although there are different ways to express a rise or fall, what appears to matter most (at least within the design of the experiments presented here) is the ending F0 target; alternative statistical parameterizations in terms of excursion, slope, or TCoG in practice merely served to highlight the role of ending F0. The results of Part 1 overall show a clear distinction in the interpretation of rising versus falling F0 contours, a finding which is also in line with the results of imitation paradigms showing that naïve speakers show a particularly robust distinction between falling and rising intonational tunes but that further distinctions within these broad classes (e.g., L*H-H% versus H*H-H% or H*L-L% versus H*L-H%) are less systematic across speakers (Cole et al., 2023).

The findings from Part 2 showed that participants responded similarly to RFR-shaped tunes in offline interpretation (in the context of SI), but that they respond differently to these tunes in online processing. Specifically, RFR-shaped tunes showed higher SI rates than falls (which used the same set of pitch accents), but only H*L-H% showed additional facilitation of higher alternatives while L*+HL-H% instead showed less facilitation of higher alternatives—there were no clear patterns for any of the falls. The two sets of results seem to paint conflicting pictures for a clear distinction between RFR-shaped tunes if a robust three-way categorical distinction is posited. One way to reconcile the two sets of results was proposed in terms of pitch range, which was originally proposed by Hirschberg & Ward (1992) to differentiate UNCERTAINTY from INCREDULITY in the context of RFR. In the context of the inference task presented in Part 2, SI was most likely with L*+HL-H% (i.e., the RFR with the largest pitch range) which may come as a result of increased arousal/emotional activation of the speaker serving as a cue to the listener that an inference (i.e., SI) should be drawn. For the lexical decision task, it is possible that participants considered a wider range of potential inferences when SI was not explicitly probed (e.g., consideration of SI *or* incredulity as two distinct possible inferences) which may be cognitively costly, hence leading to less facilitation of the higher alternative with L*+HL-H% compared to H*L-H%, which showed

more facilitation.

Due to the reported connection between RFR and alternatives—under some accounts, higher alternatives specifically—the results of Part 2 make a broader contribution to the literature on SI. In particular, the results replicated prior work on the effect of intonation on SI computation (i.e., that RFR makes SI more likely; de Marneffe & Tonhauser, 2019; Ronai & Göbel, 2024). The results also showed that the distinction of *higher* or *lower* alternative mattered in so far as RFR specifically modulated the facilitation of the higher, and not the lower, alternative at a later point in processing. Additionally, although the dual task employed in Part 2 appeared to mask this modulation in facilitation, the results did show that participant RTs to the lexical decision portion of the task were slower on trials where they computed the SI-enriched interpretation (i.e., responded Yes on the SI-probing question that followed the lexical decision judgment). This pattern is in line with the view that SI computation is costly, rather than automatic (Bott & Noveck, 2004; De Neys & Schaeken, 2007; Huang & Snedeker, 2009, among many others, though importantly c.f. Degen & Tanenhaus, 2015; Grodner et al., 2010 i.a.). When looking at the overall pattern of activation for higher and lower alternatives—abstracting away from intonation—the distinction didn't matter; furthermore, scalar alternatives did not differ from contrastive associates (i.e., focus alternatives). This lack of a distinction between scalar and focus alternatives is in line with a view that the alternative-generating process is the same for both types of alternatives (Fox & Katzir, 2011).

5.2 Limitations

The present thesis, like any other, has some limitations. A major theme of the present work is using perception and comprehension tasks in the presence of strictly controlled variation in the F0 patterns of the materials. In order to achieve this high level of control over the phonetic expression of the tunes investigated here, only a single speaker (myself) was used to record the various intonational manipulations (see Kurumada & Buxó-Lugo, 2024 for work on speaker adaptation with rising and falling intonation). While this work does not provide production data,[1] complementary

work on the nature of intonational categories and variation does exist (Cole et al., 2023; Dilley & Heffner, 2013; Pierrehumbert & Steele, 1989a; Steffman et al., 2024; Zahner-Ritter et al., 2022 among others). One might wonder whether distinctions that were similar in terms of perception here would be similar (or dissimilar) in production—potentially with the addition of discourse contexts like those used in Part 2 (see also Ronai & Göbel, 2024). Additionally, this work did not directly test the degree to which pairs of contours (particularly in Part 1) were discriminable from one another. For instance, although Part 1 showed that the proportion of TELLING responses were at floor for both L*H-H% (in Exp. 2c) and L*+HH-H% (in Exp. 4), it is unclear whether participants would be able to discriminate between the two if presented in a discrimination task. Further, it is unclear whether the two contours would be discriminable on **acoustic** grounds but not discriminable on **pragmatic** grounds: participants may be able to determine that the contours are two different acoustic signals but are unable to determine that they convey different meanings. Whether imitations of these two contours would be any different is similarly unexplored here.[2]

Although a great amount of phonetic variation was explored in this thesis, the focus was on the rather broad macro-variety of MAE; it was not a goal to relate the variation in the materials to potentially socially meaningful or dialect/variety-specific sociophonetic variation (e.g., Burdin et al., 2018; Conner, 2020; Holliday, 2021) within MAE or across other varieties. For instance, plateaus and other contours ending in mid-pitch in MAE were shown in Part 1 to be at-chance in terms of the proportion of TELLING responses, but these contours may very well be robustly interpreted as questions in varieties like Glaswegian English, where the analogue of "question intonation" is characterized by phrase-final mid-level pitch (Vizcaíno Ortega, 2002). Relatedly, although reference is made to sociophonetic work on uptalk (Warren, 2016) when describing the "fundamental parameters" by which rises and falls can vary, the present work did not seek to

[1]Although, see Appendix C.2 for what amounts to a rather exhaustive single-participant production study for falls and RFR-shaped tunes that differ in pitch accent. This appendix goes into detail about the variation in intonational form for these tunes across a variety of different metrical structures.

[2]Incidentally, there is ongoing production work (focusing less on interpretation) using imitation paradigms which make use of rises that differ in the shape of the trajectory from the low valley to the high ending F0 target. Sostarics & Cole (2024) provides a preliminary descriptive analysis of the rising contours that participants produced while Iskarous et al. (accepted) provides a more in-depth analysis that relates more broadly to the status of the phrase accent in MAE—a controversial area of intonational phonology in MAE.

disentangle uptalk uses of rising intonation from non-uptalk uses of rising intonation.[3]

With regard to the goals and contributions of this thesis, the major goal was to examine **phonological** contrasts in MAE by way of the meaning distinctions different tunes have been posited to convey. That is, it was not a goal to propose new pragmatic characterizations of rising declaratives or RFR, nor was it a goal to fully adjudicate between existing accounts of either phenomenon. For Part 2 in particular, although the results may be more compatible with some accounts of the meaning contribution of RFR and less compatible with other accounts, future work on the "meaning side" will need to provide more insight on this front.[4] That said, the results presented in this thesis do provide empirical data related to ongoing questions of interest to practitioners in experimental pragmatics, especially those concerned with the processing of scalar alternatives. The results more broadly provide a wealth of empirical data that future accounts can take into consideration.

Lastly, the present work has focused on native speakers/listeners of MAE, but the inventory of intonational patterns and the meaning distinctions they convey are not universal; accordingly, there may be systematic variation among bilingual or L2 listeners or variation attributable to individual differences (Bishop, 2016; Patel et al., 2019; Yu & Zellou, 2019) that was not explored here. However, with regard to the role of individual differences in particular, it should be noted that such omission was not for lack of trying: All participants in all experiments responded to the communication and social skills sub-scales of the Autism-Quotient (AQ) questionnaire (Baron-Cohen et al., 2001) as well as the UCLA perceived loneliness scale (Russell, 1996)[5] but exploratory analyses of both factors showed spurious results that were inconsistent across experiments. For instance,

[3]However, if an "uptalk interpretation" was available to participants, and if it were somehow distinct from the ASKING/TELLING options that were provided, one would expect such an interpretation to (1) fall under the umbrella of "interfering" interpretations that were described in that chapter and (2) be mentioned in the free-text responses in the final experiment of that chapter (notably, uptalk was never mentioned in any of the responses). One conjecture is that metalinguistic stereotypes in popular culture surrounding uptalk tend to be linked to women but the experiment used a man's voice, hence participants may not have thought to discuss uptalk. See Stecker (2023) for a more targeted investigation of sociolinguistic expectations and uptalk.

[4]I would, however, urge future experimental pragmatic work to be more descriptive of the acoustic materials used given that pitch range (rather than categorical pitch accent) appeared to capture the patterns across the two sets of results in the results described in Chapter 4.

[5]The loneliness scale questionnaire has been used in social neuroscience research to relate higher levels of perceived loneliness with greater difficulty in engaging with interpersonal relationships (S. Cacioppo et al., 2014), reduced brain mass in regions responsible for social behavior (J. T. Cacioppo et al., 2009; Nakagawa et al., 2015), and reduced cognitive performance (Hawkley & Cacioppo, 2010). J. T. Cacioppo & Cacioppo (2018) describe that despite

higher AQ scores were associated with slightly slower overall reaction times in one experiment but slightly faster reaction times in another. My speculation is that the ASKING/TELLING distinction in Part 1 is perhaps too robust or too simple of a task to be sensitive to meaningful variation in these factors and that the design of the experiments in Part 2 was not statistically powered enough to allow a meaningful analysis of these uncontrolled exploratory factors. Because these results were not the primary focus of this thesis, a full chapter of null results[6] was omitted from the current writing; however, future work may decide to more intently investigate these factors.

5.3 Final words

The issue of categoricity and gradience in intonational form and function is not new and disentangling the two is not easy. Yet there appears to be less converging evidence than one might expect if it were the case that strong divisions existed. Future work tackling distinctions in the tunes investigated here may opt for different experimental designs (e.g., to better disentangle F0 excursion from slope as described in Chapter 3) or opt to target different meaning dimensions (e.g., surprisal as in Gussenhoven & Rietveld, 2000), and indeed such investigations may find different patterns than what was found here. Nonetheless, it should be kept in mind that even seemingly discrete distinctions may reflect extreme points of an underlyingly gradient parameter.[7] For instance, with reference to the UNCERTAINTY and INCREDULITY readings of RFR (Hirschberg & Ward, 1992), Calhoun (2004, pp. 84–85) and Ladd (2022, p. 252) note that even though these two readings can be treated as categorically distinct from one another, the distinction can arise without the need for

loneliness being a motivator for people to seek out social interaction, loneliness also counterintuitively "arouses a conflicting motivation to avoid others [...] for self-preservation" hence reducing pro-social behavior. The idea here was that engagement with pragmatic reasoning is inherently social in nature, and so a participant who displays higher scores on the loneliness scale may be less likely to engage in pragmatic reasoning in this study or be less sensitive to the distinctions conveyed by intonation.

[6]To be clear though, this is not to say that there did not exist variation across individuals. For instance, in the context of RT, there will always be some people who are a little bit faster than average and some people who are a little bit slower than average. In fact, this distribution of by-participant deviations is precisely what is accounted for with by-participant random intercept/slopes. What is at issue here is that these deviations do not appear to systematically relate to measures like the AQ questionnaire.

[7]A broader question is whether such gradient phonetic parameters are biologically driven and universal, e.g. following the biological codes described by Gussenhoven (2004), and to what degree a repertoire of such gradient parameters is culturally mediated or language-specific.

two separate tunes (in line with the pitch range proposal of Hirschberg & Ward, 1992).[8]

To conclude, work on rising declaratives and RFR have largely been two disjoint areas of research in intonational meaning, but both lines of work show a common uncertainty with regard to which distinctions in intonational form matter for distinctions in meaning. The goal of this thesis was to reconnect these areas of research with a rigorous investigation of intonational **form**, tackling the question of the relevance of between- and within-category variation. Although it is uncontroversial that there are robust differences among different intonational patterns in MAE like rising, falling, and RFR intonation, the results of the present work, taken together, suggest that strong category-level distinctions within these broad classes are likely not as robust. Whether the distinctions within these tune classes are secondary and require just the right discourse context to eke out the systematic difference between, say, H*L-H% and L+H*L-H% remains possible—but the results presented here suggest this may be unlikely. A more fruitful characterization of the results presented here may be to appeal to meaningful phonetic gradience, echoing the suggestion from Ladd (2022, p. 254) that although we need some way to describe different nuances related to phonetic variation, "we have to be content to describe those nuances in gradient, statistical, phonetic terms" rather than convenient yet strict categorical labels like L*H-H% and H*H-H% or H*L-H%, L+H*L-H%, and L*+HL-H%.

[8]Related to this, it should be noted that the expression of INCREDULITY, or perhaps SURPRISAL or MIRATIVITY (Cruschina, 2021) more broadly, does not appear to be restricted to RFR-shaped tunes specifically. Rises with large magnitudes (like those used in Part 1) can express speaker incredulity, as shown by the number of free-text responses offering *surprisal* as a salient nuance (i.e., the incredulous rising declaratives of Goodhue, 2024; e.g., "We're having a *BABY?*"). A more proment high pitch accent (Rett & Sturman, 2020) may similarly express that information is surprising (or should be surprising to the hearer, e.g., "We're having a *BABY!*"). See Dessì Schmid et al. (2025) for a related typology of mirative expressions in terms of speaker- and hearer-orientedness. While the discourse context likely mediates the choice of tune (rise, fall, or RFR) and whether expressing surprisal is appropriate, it seems likely that the larger pitch range proposed for the incredulity reading of RFR would generalize for other tunes. Further work similar to Liberman & Pierrehumbert (1984) and Gussenhoven & Rietveld (2000) would likely be a fruitful avenue. There is also ongoing work exploring surprisal with rising and falling intonation building on the results of this thesis (Stanhope et al., accepted).

REFERENCES

Aparicio, H., & Ronai, E. (2023). Scalar implicature rates vary within and across adjectival scales. *Semantics and Linguistic Theory*, 110–130. https://doi.org/10.3765/t7t8pn98

Arnhold, A., Braun, B., & Romero, M. (2021). Aren't prosody and syntax marking bias in questions? *Language and Speech*, *64*(1), 141–180. https://doi.org/10.1177/0023830920914315

Arvaniti, A. (2019). Crosslinguistic variation, phonetic variability, and the formation of categories in intonation. *Proceedings of the International Congress of the Phonetic Sciences*.

Arvaniti, A., Gryllia, S., Zhang, C., & Marcoux, K. (2022). Disentangling emphasis from pragmatic contrastivity in the english h*˜ l+ h* contrast. In J. Barnes & S. Shattuck-Hufnagle (Eds.), *Proceedings of speech prosody* (pp. 23–26, Vol. 2022). MIT Press.

Arvaniti, A., Katsika, A., & Hu, N. (2024). Variability, overlap, and cue trading in intonation. *Language*, *100*(2), 265–307. https://doi.org/https://doi.org/10.1353/lan.2024.a929737

Baddeley, A. (2003). Working memory and language: An overview [ASHA 2002]. *Journal of Communication Disorders*, *36*(3), 189–208. https://doi.org/10.1016/S0021-9924(03)00019-4

Balota, D. A., Yap, M. J., Hutchison, K. A., Cortese, M. J., Kessler, B., Loftis, B., Neely, J. H., Nelson, D. L., Simpson, G. B., & Treiman, R. (2007). The english lexicon project. *Behavior research methods*, *39*, 445–459. https://doi.org/10.3758/bf03193014

Barnes, J., Brugos, A., Veilleux, N., & Shattuck-Hufnagel, S. (2021). On (and off) ramps in intonational phonology: Rises, falls, and the tonal center of gravity. *Journal of Phonetics*, *85*. https://doi.org/https://doi.org/10.1016/j.wocn.2020.101020

Barnes, J., Veilleux, N., Brugos, A., & Shattuck-Hufnagel, S. (2012). Tonal center of gravity: A global approach to tonal implementation in a level-based intonational phonology. *Laboratory Phonology*, *3*(2), 337–383.

Baron-Cohen, S., Wheelwright, S., Skinner, R., Martin, J., & Clubley, E. (2001). The autism-spectrum quotient (aq): Evidence from asperger syndrome/high-functioning autism, malesand females, scientists and mathematicians. *Journal of autism and developmental disorders*, *31*, 5–17.

Bartels, C. (1997). *Towards a compositional interpretation of english statement and question intonation* [Doctoral dissertation, University of Massachusetts Amherst].

Baumann, S., Grice, M., & Benzmüller, R. (2000). Gtobi–a phonological system for the transcription of german intonation. *Prosody*, 21–28.

Baumann, S., & Riester, A. (2013). Coreference, lexical givenness and prosody in german. *Lingua*, *136*, 16–37.

Beckman, M. E., & Pierrehumbert, J. B. (1986). Intonational structure in japanese and english. *Phonology*, *3*, 255–309. https://doi.org/10.1017/S095267570000066X

Bernstein, S. (1912). Démonstration du théoreème de weierstrass fondeée sur le calcul des probabiliteés. *Communications of the Kharkov Mathematical Society*, *13*(1), 1–2.

Bishop, J. (2016). Individual differences in top-down and bottom-up prominence perception. *Proc. Speech Prosody 2016*, *2016*, 668–672. https://doi.org/10.21437/SpeechProsody.2016-137

Bishop, J., Kuo, G., & Kim, B. (2020). Phonology, phonetics, and signal-extrinsic factors in the perception of prosodic prominence: Evidence from rapid prosody transcription. *Journal of Phonetics*, *82*, 100977. https://doi.org/10.1016/j.wocn.2020.100977

Blyth, C. R. (1972). On simpson's paradox and the sure-thing principle. *Journal of the American Statistical Association*, *67*(338), 364–366. https://doi.org/10.1080/01621459.1972.10482387

Boersma, P., & Weenink, D. (2020a). Praat: Doing phonetics by computer [computer program]. version 6.2,14.

Boersma, P., & Weenink, D. (2020b). Praat: Doing phonetics by computer [computer program]. version 6.2.14.

Bolinger, D. (1951). Intonation: Levels versus configurations. *Word*, *7*(3), 199–210.

Bolinger, D. (1978). Intonation across languages. *Universals of human language*, *2*.

Bott, L., & Chemla, E. (2016). Shared and distinct mechanisms in deriving linguistic enrichment. *Journal of Memory and Language*, *91*, 117–140.

Bott, L., & Frisson, S. (2022). Salient alternatives facilitate implicatures. *Plos one*, *17*(3), e0265781. https://doi.org/10.1371/journal.pone.0265781

Bott, L., & Noveck, I. A. (2004). Some utterances are underinformative: The onset and time course of scalar inferences. *Journal of memory and language*, *51*(3), 437–457.

Braun, B. (2006). Phonetics and phonology of thematic contrast in german. *Language and Speech*, *49*(4), 451–493. https://doi.org/10.1177/00238309060490040201

Braun, B., & Tagliapietra, L. (2010). The role of contrastive intonation contours in the retrieval of contextual alternatives. *Language and Cognitive Processes*, *25*(7-9), 1024–1043.

Buccola, B., & Goodhue, D. (2023). The effect of intonation on scalar and ignorance inferences. *Proceedings of the Chicago Linguistic Society*, *59*, 1–12.

Buccola, B., & Haida, A. (2019). Obligatory irrelevance and the computation of ignorance inferences. *Journal of Semantics*, *36*(4), 583–616.

Burdin, R. S., Holliday, N., & Reed, P. (2018). Rising Above the Standard: Variation in L+H* contour use across 5 varieties of American English. *Proc. Speech Prosody 2018*, 354–358. https://doi.org/10.21437/SpeechProsody.2018-72

Burdin, R. S., & Tyler, J. (2018). Rises inform, and plateaus remind: Exploring the epistemic meanings of "list intonation" in american english. *Journal of Pragmatics*, *136*, 97–114.

Büring, D. (2003). On d-trees, beans, and b-accents. *Linguistics and Philosophy*, *26*(5), 511–545. https://doi.org/10.1023/a:1025887707652

Büring, D. (2016). *Intonation and meaning*. Oxford University Press.

Bürkner, P.-C. (2021). Bayesian item response modeling in R with brms and Stan [R package version 2.21.6]. *Journal of Statistical Software*, *100*(5), 1–54. https://doi.org/10.18637/jss.v100.i05

Cacioppo, J. T., & Cacioppo, S. (2018). Loneliness in the modern age: An evolutionary theory of loneliness (etl). In *Advances in experimental social psychology* (pp. 127–197, Vol. 58). Elsevier.

Cacioppo, J. T., Norris, C. J., Decety, J., Monteleone, G., & Nusbaum, H. (2009). In the eye of the beholder: Individual differences in perceived social isolation predict regional brain activation to social stimuli. *Journal of cognitive neuroscience*, *21*(1), 83–92.

Cacioppo, S., Capitanio, J. P., & Cacioppo, J. T. (2014). Toward a neurology of loneliness. *Psychological bulletin*, *140*(6), 1464.

Calhoun, S. (2004). Phonetic dimensions of intonational categories - the case of l+h* and h*. *Speech Prosody 2004*, 103–106. https://doi.org/10.21437/SpeechProsody.2004-24

Calhoun, S. (2006). *Information structure and the prosodic structure of english: A probabilistic relationship* [Doctoral dissertation, The University of Edinburgh].

Calhoun, S. (2012). The theme/rheme distinction: Accent type or relative prominence? *Journal of Phonetics*, *40*(2), 329–349. https://doi.org/https://doi.org/10.1016/j.wocn.2011.12.001

Chevallier, C., Noveck, I. A., Nazir, T., Bott, L., Lanzetti, V., & Sperber, D. (2008). Making disjunctions exclusive. *Quarterly Journal of Experimental Psychology*, *61*(11), 1741–1760.

Chierchia, G. (2004). Scalar implicatures, polarity phenomena, and the syntax/pragmatics interface. In *Structures and beyond* (pp. 39–103). Oxford University Press.

Chodroff, E. R., & Cole, J. (2018). Information structure, affect, and prenuclear prominence in american english. *Proceedings of INTERSPEECH 2018*, 1848–1852.

Chodroff, E. R., & Cole, J. (2019). The phonological and phonetic encoding of information status in american english nuclear accents. *International Congress of Phonetic Sciences*, 187.

Cole, J. (2015). Prosody in context: a review. *Language, Cognition and Neuroscience*, *30*(1-2). https://doi.org/10.1080/23273798.2014.963130

Cole, J., & Chodroff, E. (2020). Prosodic encoding of information structure in nuclear and prenuclear positions in American English. *The Journal of the Acoustical Society of America*, *148*(4_Supplement), 2725–2725. https://doi.org/10.1121/1.5147567

Cole, J., & Steffman, J. (2021). The primacy of the rising/non-rising dichotomy in american english intonational tunes. *Proc. 1st International Conference on Tone and Intonation (TAI)*, 122–126.

Cole, J., Steffman, J., Shattuck-Hufnagel, S., & Tilsen, S. (2023). Hierarchical distinctions in the production and perception of nuclear tunes in american english. *Laboratory Phonology*, *14*(1). https://doi.org/10.16995/labphon.9437

Conner, T. (2020). Questioning Questions: The Illusion of Variation in African American English Polar Question Intonation. *Proc. Speech Prosody 2020*, 220–224. https://doi.org/10.21437/SpeechProsody.2020-45

Constant, N. (2012). English rise-fall-rise: A study in the semantics and pragmatics of intonation. *Linguistics and Philosophy*, *35*(5), 407–442. https://doi.org/10.1007/s10988-012-9121-1

Constant, N. (2014). *Contrastive topic: Meanings and realizations* [Doctoral dissertation, University of Massachussets Amherst].

Cruschina, S. (2021). The greater the contrast, the greater the potential: On the effects of focus in syntax. *Glossa: a journal of general linguistics*, *6*(1).

Crystal, D. (1969). *Prosodic systems and intonation in english*. Cambridge University Press.

De Carvalho, A., Reboul, A. C., der Henst, V., Cheylus, A., & Nazir, T. (2016). Scalar implicatures: The psychological reality of scales. *Frontiers in psychology*, *7*, 203305.

De Neys, W., & Schaeken, W. (2007). When people are more logical under cognitive load: Dual task impact on scalar implicature. *Experimental psychology*, *54*(2), 128–133.

de Faget de Casteljau, P. (1986). *Mathématiques et cao 2: Formes à pôles*. Hermès.

Degen, J., & Tanenhaus, M. K. (2015). Processing scalar implicature: A constraint-based approach. *Cognitive science*, *39*(4), 667–710.

de Marneffe, M.-C., & Tonhauser, J. (2019). Inferring meaning from indirect answers to polar questions: The contribution of the rise-fall-rise contour. In *Questions in discourse* (pp. 132–163). Brill. https://doi.org/10.1163/9789004378322_006

Dessì Schmid, S., Momma, L., & Wiesinger, E. (2025). Mirativity in romance: Speaker-oriented vs. hearer-oriented expressions of unexpectedness. In S. R. Rosique (Ed.), *Expressing surprise at the crossroads* (pp. 229–246). De Gruyter Mouton. https://doi.org/10.1515/9783111386683-010

Dilley, L. C., & Brown, M. (2007). Effects of pitch range variation on f0 extrema in an imitation task. *Journal of Phonetics*, *35*(4), 523–551.

Dilley, L. C., & Heffner, C. C. (2013). The role of f0 alignment in distinguishing intonation categories: Evidence from american english. *Journal of Speech Sciences*, *3*(1), 3–67.

Domaneschi, F., Romero, M., & Braun, B. (2017). Bias in polar questions: Evidence from english and german production experiments. *Glossa*.

Doran, R., Ward, G., Larson, M., McNabb, Y., & Baker, R. E. (2012). A novel experimental paradigm for distinguishing between what is said and what is implicated. *Language*, *88*(1), 124–154.

Face, T. L. (2007). The role of intonational cues in the perception of declaratives and absolute interrogatives in castilian spanish. *Journal of Experimental Phonetics*, *16*, 185–225.

Farkas, D., & Bruce, K. B. (2010). On Reacting to Assertions and Polar Questions. *Journal of Semantics*, *27*(1), 81–118. https://doi.org/10.1093/jos/ffp010

Farkas, D., & Roelofsen, F. (2017). Division of labor in the interpretation of declaratives and interrogatives. *Journal of semantics*, *34*(2), 237–289.

Fletcher, J., & Harrington, J. (2001). High-rising terminals and fall-rise tunes in australian english. *Phonetica*, *58*(4), 215–229.

Fox, D., & Katzir, R. (2011). On the characterization of alternatives. *Natural language semantics*, *19*, 87–107.

Fraundorf, S. H., Watson, D. G., & Benjamin, A. S. (2010). Recognition memory reveals just how contrastive contrastive accenting really is. *Journal of memory and language*, *63*(3), 367–386.

Gabry, J., Češnovar, R., & Johnson, A. (2024). *Cmdstanr: R interface to 'cmdstan'* [R package version 0.8.1.9].

Geluykens, R. (1988). On the myth of rising intonation in polar questions. *Journal of Pragmatics*, *12*(4), 467–485. https://doi.org/https://doi.org/10.1016/0378-2166(88)90006-9

Göbel, A. (2019). Additives pitching in: L*+H signals ordered focus alternatives. *Semantics and Linguistic Theory*, *29*, 279–299. https://doi.org/10.3765/salt.v29i0.4612

Göbel, A., & Wagner, M. (2023a). On a concessive reading of the rise-fall-rise contour: Contextual and semantic factors. *Experiments in Linguistic Meaning*, *2*, 83–94.

Göbel, A., & Wagner, M. (2023b). A pitch accent beyond contrastive focus marking: Experimental evidence from auditory rating. *Proceedings of Sinn und Bedeutung*, *27*, 240–252.

Goldsmith, J. (1976). *Autosegmental phonology* [Doctoral dissertation, Indiana University].

Goldsmith, J. (1990). Autosegmental and metrical phonology. *Basil and Blackwell Ltd*.

Goodhue, D. (2022). All focus is contrastive: On polarity (verum) focus, answer focus, contrastive focus and givenness. *Journal of Semantics*, *39*(1), 117–158. https://doi.org/10.1093/jos/ffab018

Goodhue, D. (2024). Everything that rises must converge: Toward a unified account of inquisitive and assertive rising declaratives [Forthcoming, lingbuzz/006493]. In A. Benz, D. Goodhue, M. Krifka, T. Trinh, & K. Yatsushiro (Eds.), *Biased questions: Experimental results & theoretical modelling*. Language Science Press.

Goodhue, D., Harrison, L., Su, Y. C., & Wagner, M. (2016). Toward a bestiary of english intonational contours. *The Proceedings of the North East Linguistics Society (NELS)*, *46*, 311–320.

Gotzner, N. (2019). The role of focus intonation in implicature computation: A comparison with only and also. *Natural Language Semantics*, *27*(3), 189–226.

Gotzner, N., & Romoli, J. (2022). Meaning and Alternatives. *Annual Review of Linguistics*, *8*, 213–234. https://doi.org/10.1146/annurev-linguistics-031220-012013

Gotzner, N., Solt, S., & Benz, A. (2018). Scalar diversity, negative strengthening, and adjectival semantics. *Frontiers in Psychology*, *9*(SEP), 1–13. https://doi.org/10.3389/fpsyg.2018.01659

Gotzner, N., & Spalek, K. (2017). Role of contrastive and noncontrastive associates in the interpretation of focus particles. *Discourse Processes*, *54*(8), 638–654.

Gotzner, N., & Spalek, K. (2019). The life and times of focus alternatives: Tracing the activation of alternatives to a focused constituent in language comprehension. *Language and Linguistics Compass*, *13*(2), e12310.

Gotzner, N., Spalek, K., & Wartenburger, I. (2013). How pitch accents and focus particles affect the recognition of contextual alternatives. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *35*.

Gotzner, N., Wartenburger, I., & Spalek, K. (2016). The impact of focus particles on the recognition and rejection of contrastive alternatives. *Language and Cognition*, *8*(1), 59–95.

Grice, H. P. (1975). Logic and conversation. In *Speech acts* (pp. 41–58). Brill.

Grice, M., Reyelt, M., Benzmuller, R., Mayer, J., & Batliner, A. (1996). Consistency in transcription and labelling of german intonation with gtobi. *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP '96*, *3*, 1716–1719 vol.3. https://doi.org/10.1109/ICSLP.1996.607958

Grodner, D. J., Klein, N. M., Carbary, K. M., & Tanenhaus, M. K. (2010). "some," and possibly all, scalar inferences are not delayed: Evidence for immediate pragmatic enrichment. *Cognition*, *116*(1), 42–55.

Groenendijk, J., & Roelofsen, F. (2009). Inquisitive Semantics and Pragmatics. *Workshop on Language, Communication and Rational Agency*.

Gunlogson, C. (2001, January). *True to form: Rising and falling declaratives as questions in english* [Doctoral dissertation, University of California Santa Cruz].

Gunlogson, C. (2008). A question of commitment. *Belgian Journal of Linguistics*, *22*(1), 101–136. https://doi.org/10.1075/bjl.22.06gun

Gussenhhoven, C., & Chen, A. (2020, December). *The oxford handbook of language prosody*. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780198832232.001.0001

Gussenhoven, C. (1984, January). *On the grammar and semantics of sentence accents* (Vol. 16). Foris. https://doi.org/10.1515/9783110859263

Gussenhoven, C. (2004). *The phonology of tone and intonation*. Cambridge University Press.

Gussenhoven, C. (2016). Analysis of intonation: The case of MAE-ToBI. *Laboratory Phonology*, *7*(1), 1–35. https://doi.org/10.5334/labphon.30

Gussenhoven, C., & Rietveld, T. (2000). The behavior of h* and l* under variations in pitch range in dutch rising contours. *Language and speech*, *43*(2), 183–203.

Gussenhoven, C., & van de Ven, M. (2020). Categorical perception of lexical tone contrasts and gradient perception of the statement–question intonation contrast in zhumadian mandarin. *Language and Cognition*, *12*(4), 614–648. https://doi.org/10.1017/langcog.2020.14

Halliday, M. (1967). *Intonation and grammar in british english*. Mouton.

Hawkley, L. C., & Cacioppo, J. T. (2010). Loneliness Matters: A Theoretical and Empirical Review of Consequences and Mechanisms. *Annals of Behavioral Medicine*, *40*(2), 218–227. https://doi.org/10.1007/s12160-010-9210-8

Hillenbrand, J., Getty, L. A., Clark, M. J., & Wheeler, K. (1995). Acoustic characteristics of American English vowels. *The Journal of the Acoustical Society of America*, *97*(5), 3099–3111. https://doi.org/10.1121/1.411872

Hirschberg, J. (2017). Pragmatics and prosody. In *The oxford handbook of pragmatics* (pp. 532–549). Oxford University Press.

Hirschberg, J., & Ward, G. (1992). The influence of pitch range, duration, amplitude and spectral features on the interpretation of the rise-fall-rise intonation contour in english. *Journal of phonetics*, *20*(2), 241–251. https://doi.org/10.1016/S0095-4470(19)30625-4

Hirschberg, J., & Ward, G. (1995). The interpretation of the high-rise question contour in english. *Journal of Pragmatics*, *24*(4), 407–412.

Hirst, D., & Di Cristo, A. (1998). A survey of intonation systems. *Intonation systems: A survey of twenty languages*, *144*, 152–166.

Hobbs, J. R. (1990, January). The pierrehumbert-hirschberg theory of intonational meaning made simple: Comments on pierrehumbert and hirschberg. In P. Cohen, J. Morgan, & M. Pollack (Eds.), *Intentions in communication* (pp. 313–323). Cambridge Massachusetts: MIT Press.

Hogan, J., & Adams, N. M. (2023). On averaging ROC curves [Survey Certification]. *Transactions on Machine Learning Research*.

Holliday, N. (2021). Prosody and sociolinguistic variation in american englishes. *Annual review of linguistics*, *7*(1), 55–68.

Holliday, N., & Villarreal, D. (2020). Intonational variation and incrementality in listener judgments of ethnicity. *Laboratory Phonology: Journal of the Association for Laboratory Phonology*, *11*(1), 3.

Horn, L. R. (1972). *On the semantic properties of logical operators in english* [Doctoral dissertation, University of California, Los Angeles].

Huang, Y. T., & Snedeker, J. (2009). Online interpretation of scalar quantifiers: Insight into the semantics–pragmatics interface. *Cognitive psychology*, *58*(3), 376–415.

Husband, E. M., & Ferreira, F. (2016). The role of selection in the comprehension of focus alternatives. *Language, Cognition and Neuroscience*, *31*(2), 217–235.

Im, S., Cole, J., & Baumann, S. (2023). Standing out in context: Prominence in the production and perception of public speech. *Laboratory Phonology*, *14*. https://doi.org/10.16995/labphon.6417

Iskarous, K., bf Sostarics, T., & Cole, J. (accepted). The dynamical structure of the nuclear tune: The phrase accent. *Proc. 3rd International Conference on Tone and Intonation (TAI)*.

Iskarous, K., Cole, J., & Steffman, J. (2024). A minimal dynamical model of intonation: Tone contrast, alignment, and scaling of american english pitch accents as emergent properties. *Journal of Phonetics*, *104*, 101309. https://doi.org/https://doi.org/10.1016/j.wocn.2024.101309

Jackendoff, R. (1972). Semantic interpretation in generative grammar.

Jacquemot, C., & Scott, S. K. (2006). What is the relationship between phonological short-term memory and speech processing? *Trends in Cognitive Sciences*, *10*(11), 480–486. https://doi.org/https://doi.org/10.1016/j.tics.2006.09.002

Jeong, S. (2018). Intonation and Sentence Type Conventions: Two Types of Rising Declaratives. *Journal of Semantics*, *35*(2), 305–356. https://doi.org/10.1093/semant/ffy001

Jun, S.-A. (2022). The tobi transcription system: Conventions, strengths, and challenges. In J. Barnes & S. Shattuck-Hufnagle (Eds.), *Prosodic theory and practice*. MIT Press.

Kock, N., & Gaskins, L. (2016). Simpson's paradox, moderation and the emergence of quadratic relationships in path models: An information systems illustration. *International Journal of Applied Nonlinear Science*, *2*, 200. https://doi.org/10.1504/IJANS.2016.077025

Krahmer, E., & Swerts, M. (2001). On the alleged existence of contrastive accents. *Speech Communication*, *34*(4), 391–405. https://doi.org/https://doi.org/10.1016/S0167-6393(00)00058-3

Krifka, M. (2008). Basic notions of information structure. *Acta Linguistica Hungarica (Since 2017 Acta Linguistica Academica)*, *55*(3-4), 243–276.

Kurumada, C., & Buxó-Lugo, A. (2024). Intonation adaptation to multiple talkers. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *50*, 1954–1981. https://doi.org/10.1037/xlm0001419

Lacina, R., Alexandropoulou, S., Ronai, E., & Gotzner, N. (2024). *Scalar alternative activation in implicature processing: A lexical decision study with antonyms and negation* [Preprint]. https://doi.org/https://doi.org/10.31234/osf.io/r3q79

Lacina, R., & Gotzner, N. (2024). Exploring scalar diversity through priming: A lexical decision study with adjectives. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *46*.

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

Ladd, D. R. (2015). The american four-level analysis of intonation contours. *Historiographia linguistica*, *42*(1).

Ladd, D. R. (2022). The trouble with ToBI. In J. Barnes & S. Shattuck-Hufnagle (Eds.), *Prosodic theory and practice* (pp. 247–259). MIT Press.

Ladd, D. R., & Morton, R. (1997). The perception of intonational emphasis: Continuous or categorical? *Journal of Phonetics*, *25*(3), 313–342.

Ladd, D. R., & Schepman, A. (2003). "sagging transitions" between high pitch accents in english: Experimental evidence. *Journal of phonetics*, *31*(1), 81–112.

Ladd, D. R., Silverman, K., Tolkmitt, F., Bergmann, G., & Scherer, K. (1985). Evidence for the independent function of intonation contour type, voice quality, and f0 range in signaling speaker affect. *Journal of The Acoustical Society of America - J ACOUST SOC AMER*, *78*(2), 435–444. https://doi.org/10.1121/1.392466

Leung, K. K. W., Jongman, A., Wang, Y., & Sereno, J. A. (2016). Acoustic characteristics of clearly spoken English tense and lax vowelsa). *The Journal of the Acoustical Society of America*, *140*(1), 45–58. https://doi.org/10.1121/1.4954737

Levinson, S. C., & Torreira, F. (2015). Timing in turn-taking and its implications for processing models of language. *Frontiers in Psychology*, *6*. https://doi.org/10.3389/fpsyg.2015.00731

Liberman, M., & Pierrehumbert, J. B. (1984, January). Intonational invariance under changes in pitch range and length. In *Language sound structure: Studies in phonology presented to morris halle* (pp. 157–233, Vol. 157). MIT Press.

Liberman, M., & Sag, I. (1974). Prosodic form and discourse function. *Chicago Linguistics Society*, *10*, 416–427.

Lorenzen, J., Roessig, S., & Baumann, S. (2023). Redundancy and individual variability in the prosodic marking of information status in german. *Proceedings of the 20th International Congress of Phonetic Sciences (ICPhS)*.

Malamud, S. A., & Stephenson, T. (2015). Three ways to avoid commitments: Declarative force modifiers in the conversational scoreboard. *Journal of Semantics*, *32*(2), 275–311. https://doi.org/10.1093/jos/ffu002

Marty, P., Romoli, J., Sudo, Y., & Breheny, R. (2024). Implicature priming, salience, and context adaptation. *Cognition*, *244*, 105667. https://doi.org/https://doi.org/10.1016/j.cognition.2023.105667

Mazzarella, D., Reinecke, R., Noveck, I., & Mercier, H. (2018). Saying, presupposing and implicating: How pragmatics modulates commitment. *Journal of Pragmatics*, *133*, 15–27.

McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., & Sonderegger, M. (2017a). Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi. *Proc. Interspeech 2017*, 498–502. https://doi.org/10.21437/Interspeech.2017-1386

McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., & Sonderegger, M. (2017b). Montreal forced aligner: Trainable text-speech alignment using kaldi. *Interspeech*, *2017*, 498–502.

Muxica, C., & Harris, J. A. (2025). Constructing alternatives: Evidence for the early availability of contextually relevant focus alternatives. In *Alternatives in grammar and cognition* (pp. 75–118). Springer.

Nakagawa, S., Takeuchi, H., Taki, Y., Nouchi, R., Sekiguchi, A., Kotozaki, Y., Miyauchi, C. M., Iizuka, K., Yokoyama, R., Shinada, T., et al. (2015). White matter structures associated with loneliness in young adults. *Scientific Reports*, *5*(1), 1–11.

Niebuhr, O., Skarnitzl, R., & Tylečková, L. (2018). The acoustic fingerprint of a charismatic voice - Initial evidence from correlations between long-term spectral features and listener ratings. *Proc. Speech Prosody 2018*, 359–363. https://doi.org/10.21437/SpeechProsody.2018-73

Nilsenová, M. (2006). *Rises and falls. studies in the semantics and pragmatics of intonation* [Doctoral dissertation, University of Amsterdam].

Nolan, F. (2022, February). The rise and fall of the british school of intonation analysis. In *Prosodic theory and practice*. The MIT Press. https://doi.org/10.7551/mitpress/10413.003.0012

Olsen, A. (2018). *Bezier: Toolkit for bezier curves and splines* [R package version 1.1.2].

Orrico, R., Gryllia, S., Kim, J., & Arvaniti, A. (2025). Individual variability and the h* l + h* contrast in english. *Language and Cognition*, *17*, e9. https://doi.org/10.1017/langcog.2024.62

Patel, S. P., Kim, J. H., Larson, C. R., & Losh, M. (2019). Mechanisms of voice control related to prosody in autism spectrum disorder and first-degree relatives. *Autism Research*. https://doi.org/10.1002/aur.2156

Peirce, J. W. (2007). Psychopy–psychophysics software in python. *Journal of neuroscience methods*, *162*(1-2), 8–13. https://doi.org/10.1016/j.jneumeth.2006.11.017

Petrone, C., & Niebuhr, O. (2014). On the intonation of german intonation questions: The role of the prenuclear region. *Language and Speech*, *57*(1), 108–146.

Phillips, G. M. (2006). *Interpolation and approximation by polynomials*. Springer Science & Business Media.

Pierrehumbert, J. B. (1980). *The phonology and phonetics of english intonation* [Doctoral dissertation, Massachusetts Institute of Technology].

Pierrehumbert, J. B., & Hirschberg, J. (1990, January). The meaning of intonational contours in the interpretation of discourse. In P. Cohen, J. Morgan, & M. Pollack (Eds.), *Intentions in communication* (pp. 271–311). Cambridge Massachusetts: MIT Press.

Pierrehumbert, J. B., & Steele, S. (1989a). Categories of tonal alignment in english. *Phonetica*, *46*(4), 181–196.

Pierrehumbert, J. B., & Steele, S. (1989b). Categories of tonal alignment in english. *Phonetica*, *46*(4), 181–196.

Pitrelli, J. F., Beckman, M. E., & Hirschberg, J. (1994). Evaluation of prosodic transcription labeling reliability in the tobi framework. *Icslp*, *1*, 123–6.

Prieto, P. (2015). Intonational meaning. *Wiley Interdisciplinary Reviews: Cognitive Science*, *6*(4), 371–381.

Prince, E. (1981). Toward a taxonomy of given-new information. In P. Cole (Ed.), *Radical pragmatics* (pp. 223–255). Academic Press.

R Core Team. (2024). *R: A language and environment for statistical computing* [version 4.4.0]. R Foundation for Statistical Computing. Vienna, Austria.

Rees, A., & Bott, L. (2018). The role of alternative salience in the derivation of scalar implicatures. *Cognition*, *176*, 1–14.

Repp, S., & Seeliger, H. (2020). Prosodic prominence in polar questions and exclamatives. *Frontiers in communication*, *5*, 53.

Repp, S., & Seeliger, H. (2023). Contrast and givenness in biased declarative questions. *Proceedings of the 20th ICPhS, R. Skarnitzl and J. Volın, Eds. Prague: Guarant International*, 1543–1547.

Rett, J., & Sturman, B. (2020). Prosodically marked mirativity. *Proceedings of WCCFL*, *38*.

Roberts, C. (1996). Information structure in discourse: Toward a unified theory of formal pragmatics. *Ohio State University Working Papers in Linguistics*, *49*, 91–136.

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., & Müller, M. (2011). Proc: An open-source package for r and s+ to analyze and compare roc curves. *BMC Bioinformatics*, *12*, 77.

Roessig, S. (2021). *Categoriality and continuity in prosodic prominence (volume 10)*. Language Science Press.

Roessig, S. (2024). The inverse relation of pre-nuclear and nuclear prominences in german. *Laboratory Phonology*, *15*(1), 1–43.

Roessig, S., Mücke, D., & Grice, M. (2019). The dynamics of intonation: Categorical and continuous variation in an attractor-based model. *PloS one*, *14*(5), e0216859.

Roettger, T., Mahrt, T., & Cole, J. (2019). Mapping prosody onto meaning – the case of information structure in american english. *Language, Cognition and Neuroscience*, *34*(7), 841–860. https://doi.org/10.1080/23273798.2019.1587482

Ronai, E., & Göbel, A. (2024). Watch your tune! on the role of intonation for scalar diversity. *Glossa Psycholinguistics*, *3*(1).

Ronai, E., Sun, Y., Yu, A. C., & Xiang, M. (2019). Integration of contextual-pragmatic and phonetic information in speech perception: An eye-tracking study. *Laboratory Phonology*, *10*(1), 1–15. https://doi.org/10.5334/labphon.186

Ronai, E., & Xiang, M. (2023). Tracking the activation of scalar alternatives with semantic priming. *Experiments in Linguistic Meaning*, *2*, 229–240. https://doi.org/10.3765/elm.2.5371

Ronai, E., & Xiang, M. (2024). What could have been said? Alternatives and variability in pragmatic inferences. *Journal of Memory and Language*, *136*, 104507.

Rooth, M. (1992). A theory of focus interpretation. *Natural language semantics*, *1*(1), 75–116.

Rudin, D. (2022). Intonational Commitments. *Journal of Semantics*, *39*(2), 339–383. https://doi.org/10.1093/jos/ffac002

Russell, D. W. (1996). Ucla loneliness scale (version 3): Reliability, validity, and factor structure. *Journal of personality assessment*, *66*(1), 20–40.

Sandberg, K. (2024). *The interpretation of prosodic prominence conveying contrast and intensity* [Doctoral dissertation, Northwestern University].

Sauerland, U. (2004). Scalar implicatures in complex sentences. *Linguistics and philosophy*, *27*, 367–391.

Schad, D. J., Vasishth, S., Hohenstein, S., & Kliegl, R. (2020). How to capitalize on a priori contrasts in linear (mixed) models: A tutorial. *Journal of Memory and Language*, *110*, 104038. https://doi.org/https://doi.org/10.1016/j.jml.2019.104038

Schiefer, L., & Batliner, A. (1991). A ramble round the order effect. *Forschungsberichte des Instituts für Phonetik und Sprachliche Kommunikation der Universität München*.

Schwarz, F., Clifton Jr, C., & Frazier, L. (2007). Strengthening 'or': Effects of focus and downward entailing contexts on scalar implicatures. *University of Massachusetts Occasional Papers in Linguistics*, *33*(1), 9.

Schwarz, F., Zehr, J., Grodner, D., & Bacovcin, H. A. (2016). *Subliminal priming of alternatives does not increase implicature responses* [Poster presented at the Logic and Language in Conversation Workshop, University of Utrecht].

Seeliger, H., & Repp, S. (2023). Information-structural surprises? contrast, givenness, and (the lack of) accent shift and deaccentuation in non-assertive speech acts. *Laboratory Phonology*, *14*(1).

Sellers, K., Lotze, T., & Raim, A. (2023). *Compoissonreg: Conway-maxwell poisson (com-poisson) regression* [R package version 0.8.1].

Silverman, K. E., Beckman, M. E., Pitrelli, J. F., Ostendorf, M., Wightman, C. W., Price, P., Pierrehumbert, J. B., & Hirschberg, J. (1992). Tobi: A standard for labeling english prosody. *ICSLP*, *2*, 867–870.

Simpson, E. H. (1951). The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, *13*(2), 238–241. https://doi.org/10.1111/j.2517-6161.1951.tb00088.x

Sostarics, T. (2024). *contrastable: Contrast coding utilities in R* [R package version 1.0.2.9000]. https://doi.org/10.32614/CRAN.package.contrastable

Sostarics, T., & Cole, J. (2021). Epistemic Meaning and the LLL Tune in American English. *Proc. 1st International Conference on Tone and Intonation (TAI)*, 11–15. https://doi.org/10.21437/TAI.2021-3

Sostarics, T., & Cole, J. (2023a). Pitch Accent Variation and the Interpretation of Rising and Falling Intonation in American English. *Proc. INTERSPEECH 2023*, 97–101. https://doi.org/10.21437/Interspeech.2023-315

Sostarics, T., & Cole, J. (2023b). Testing the Locus of Speech-Act Meaning in English Intonation. In R. Skarnitzl & J. Volín (Eds.), *Proceedings of the 20th International Congress of Phonetic Sciences* (pp. 1240–1244). Guarant International.

Sostarics, T., & Cole, J. (2024). PitchMendR: A tool for the diagnosis and treatment of F0 irregularities. *Proceedings of Speech Prosody 2024*.

Sostarics, T., Ronai, E., & Cole, J. (2025). Relating Scalar Inference and Alternative Activation: A view from the Rise-Fall-Rise Tune in American English. *Proceedings of Experiments in Linguistic Meaning 3*, 383–394. https://doi.org/10.3765/elm.3.5768

Spalek, K., Gotzner, N., & Wartenburger, I. (2014). Not only the apples: Focus sensitive particles improve memory for information-structural alternatives. *Journal of Memory and Language*, *70*, 68–84.

Stalnaker, R. C. (1978). Assertion. In *Pragmatics* (pp. 315–332). Brill.

Stanhope, R., Sostarics, T., & Cole, J. (accepted). F0 correlates of perceived speaker surprise in American English: Accents vs. Edge Tones. *Proc. 3rd International Conference on Tone and Intonation (TAI)*.

Stecker, A. (2023). *Social expectations in linguistic memory* [Doctoral dissertation, Northwestern University].

Steffman, J., & Cole, J. (2024). Metrical enhancement in american english nuclear tunes. *Glossa: a journal of general linguistics*, *9*(1).

Steffman, J., Cole, J., & Shattuck-Hufnagel, S. (2024). Intonational categories and continua in american english rising nuclear tunes. *Journal of Phonetics*, *104*, 101310. https://doi.org/ https://doi.org/10.1016/j.wocn.2024.101310

Sun, C., Tian, Y., & Breheny, R. (2018). A link between local enrichment and scalar diversity. *Frontiers in Psychology*, *9*, 2092.

Swinney, D. A., Onifer, W., Prather, P., & Hirshkowitz, M. (1979). Semantic facilitation across sensory modalities in the processing of individual words and sentences. *Memory & Cognition*, *7*, 159–165.

Syrdal, A. K., & McGory, J. (2000). Inter-transcriber reliability of tobi prosodic labeling. *Sixth International Conference on Spoken Language Processing*.

Tabossi, P. (1996). Cross-modal semantic priming. *Language and cognitive processes*, *11*(6), 569–576.

Thorward, J. (2009). *The interaction of contrastive stress and grammatical context in child english speakers' interpretations of existential quantifiers* [Bachelor's Thesis]. Ohio State University.

Tilly, J., & Janetos, N. (2021). *Matchingr: Matching algorithms in r and c++* [R package version 1.3.3].

Trager, G. L., & Smith, H. L. (1957). *An outline of english structure*. American Council of Learned Societies.

Truckenbrodt, H. (2012). Semantics of intonation. In C. Maienborn, K. von Heusinger, & P. Portner (Eds.), *Semantics: An international handbook of natural language meaning* (pp. 2039–2969). Mouton de Gruyter. https://doi.org/10.1515/9783110589849-001

Txurruka, I. B. (2023). Perception of spanish declarative questions and statements by l2 spanish speakers. *Proceedings of the 20th international congress of phonetic sciences*, 1280–1284.

Van Tiel, B., Van Miltenburg, E., Zevakhina, N., & Geurts, B. (2016). Scalar diversity. *Journal of semantics*, *33*(1), 137–175. https://doi.org/10.1093/jos/ffu017

Vehtari, A., Gabry, J., Magnusson, M., Yao, Y., Bürkner, P.-C., Paananen, T., & Gelman, A. (2024). Loo: Efficient leave-one-out cross-validation and waic for bayesian models [R package version 2.8.0.9000].

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical bayesian model evaluation using leave-one-out cross-validation and waic. *Statistics and computing*, *27*, 1413–1432.

Veilleux, N., Shattuck-Hufnagel, S., & Brugos, A. (2006). Transcribing prosodic structure of spoken utterances with tobi [Retrieved from https://ocw.mit.edu]. *Massachusetts Institute of Technology: MIT OpenCourseWare. Retrieved from https://ocw. mit. edu License: Creative Commons BY-NC-SA*.

Vizcaíno Ortega, F. (2002). A preliminary analysis of yes/no questions in glasgow english. *Speech Prosody 2002*, 683–686. https://doi.org/10.21437/SpeechProsody.2002-156

Wagner, M. (2012). Contrastive topics decomposed. *Semantics and Pragmatics*, *5*, 8–1.

Wagner, M. (2020). Prosodic focus. In *The wiley blackwell companion to semantics* (pp. 1–75). John Wiley & Sons, Ltd. https://doi.org/10.1002/9781118788516.sem133

Wagner, M., McClay, E., & Mak, L. (2013). Incomplete answers and the rise-fall-rise contour. *Proceedings of the 17th Workshop on the Semantics and Pragmatics of Dialogue*, 140–149. https://doi.org/http://semdial.org/anthology/Z13-Wagner_semdial_0018.pdf

Waldon, B., & Degen, J. (2020). Symmetric alternatives and semantic uncertainty modulate scalar inference. *CogSci*.

Wang, J.-L., Chiou, J.-M., & Müller, H.-G. (2016). Functional data analysis. *Annual Review of Statistics and its application*, *3*(1), 257–295.

Ward, G., & Hirschberg, J. (1985). Implicating uncertainty: The pragmatics of fall-rise intonation. *Language*, *61*, 747–776. https://doi.org/10.2307/414489

Warren, P. (2014). Sociophonetic and prosodic influences on judgements of sentence type. *Proceedings of the 15th Australasian International Conference on Speech Science and Technology*, 185–188.

Warren, P. (2016). *Uptalk: The phenomenon of rising intonation*. Cambridge University Press.

Warren, P., & Fletcher, J. (2016). Phonetic differences between uptalk and question rises in two Antipodean English varieties. *Proc. Speech Prosody 2016*, 148–152. https://doi.org/10.21437/SpeechProsody.2016-31

Watson, D. G. (2010). The many roads to prominence: Understanding emphasis in conversation. In *Psychology of learning and motivation* (pp. 163–183, Vol. 52). Elsevier.

Watson, D. G., Tanenhaus, M. K., & Gunlogson, C. A. (2008). Interpreting pitch accents in online comprehension: H* vs. L+H. *Cognitive science*, *32*(7), 1232–1244.

Westera, M. (2017). *Exhaustivity and intonation: A unified theory* [Doctoral dissertation, University of Amsterdam]. Institute for Logic, Language; Computation, Universiteit van Amsterdam.

Westera, M. (2019). Rise-fall-rise as a marker of secondary QUDs. In *Secondary content* (pp. 376–404). Brill. https://doi.org/10.1163/9789004393127_015

Westera, M., Goodhue, D., & Gussenhoven, C. (2021). Meanings of tones and tunes. *The Oxford Handbook of Language Prosody. Oxford: Oxford University Press*.

Wilson, D., & Sperber, D. (2006). Relevance theory. In L. R. Horn & G. Ward (Eds.), *The handbook of pragmatics* (pp. 606–632). Wiley Online Library.

Wolf, L. (2014). *Degrees of assertion* [Doctoral dissertation, Ben Gurion University of the Negev, Faculty of Humanities and Social . . .].

Wolter, L. (2003). Fall-rise, topic, and speaker noncommitment. *Proceedings of Western Conference on Linguistics (WECOL)*, *14*, 322–333.

Wood, S. (2017). *Generalized additive models: An introduction with r* (2nd ed.). Chapman; Hall/CRC.

Xie, X., Buxó-Lugo, A., & Kurumada, C. (2021). Encoding and decoding of meaning through structured variability in intonational speech prosody. *Cognition*, *211*, 104619. https://doi.org/10.1016/j.cognition.2021.104619

Yu, A. C., & Zellou, G. (2019). Individual differences in language processing: Phonology. *Annual Review of Linguistics*, *5*(1), 131–150.

Zahner-Ritter, K., Einfeldt, M., Wochner, D., James, A., Dehé, N., & Braun, B. (2022). Three kinds of rising-falling contours in german wh-questions: Evidence from form and function. *Frontiers in Communication*, *7*. https://doi.org/10.3389/fcomm.2022.838955

Zondervan, A. (2010, May). *Scalar implicatures or focus: An experimental approach* [Doctoral thesis 1 (Research UU / Graduation UU)]. Utrecht University. LOT.

Appendix A

APPENDIX FOR CHAPTER 2 (RISING/FALLING INTONATION)

A.1 Sentences Used

The five sentences used in Chapter 2 are listed below.

1. Molly's from Branning
2. Gavin's on broadway
3. Megan's a grandma
4. Ryan's in Greenview
5. Joey's from Bronville

A.2 Experiment 2b Results

This section presents the results for Experiment 2b, which manipulated the duration of the second syllable. Figure A1 shows the empirical results while Table A1 shows the results of the statistical model.

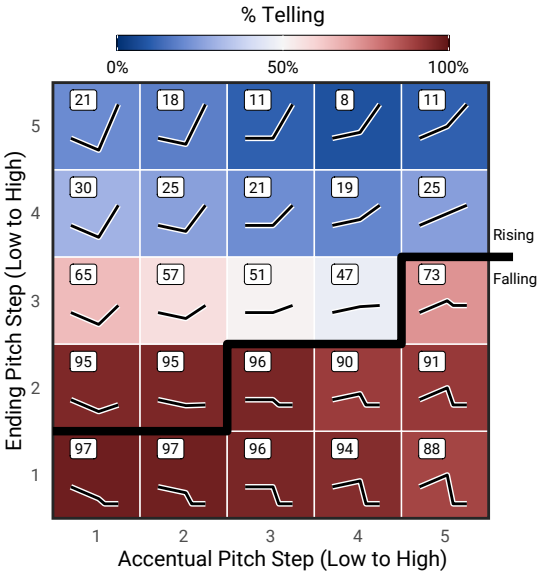


Figure A1: Exp. 2b (monotonal early falls with longer duration) aggregate proportion of TELLING responses.

Term	Estimate	Std.Error	95% CrI
Intercept	1.72	0.24	[1.25, 2.20]
AccentualPitch	−0.08	0.03	[−0.14, −0.01]
EndingPitch	−0.60	0.04	[−0.69, −0.51]
:AccentualPitch	0.01	0.01	[0.00, 0.03]

Table A1: Logistic regression model results for Experiment 2b. Estimates are shown on the log-odds scale, where higher likelihood of TELLING responses is reflected by positive values and lower likelihood of TELLING responses (=higher likelihood of ASKING responses) is reflected by negative values.

A.3 Implementation of Bitonal Accent Trajectories

Including the initial low target for a bitonal accent is straightforward; for Experiment 4 the initial low target is aligned at the beginning of the stressed syllable of *Branning* with an F0 value of 70Hz. As previously mentioned, to ensure all accentual pitch steps for Experiment 4 are rising to a higher target, the accentual pitch F0 targets are shifted up by 10Hz. To make the stimuli more natural with this additional low target, F0 is manipulated to fall from the offset of the first word of the utterance until the low F0 target. Previously, the onglide to the accentual pitch target started at the offset of the word preceding the nuclear-accented word (e.g., *from* in *Molly's from Branning*).

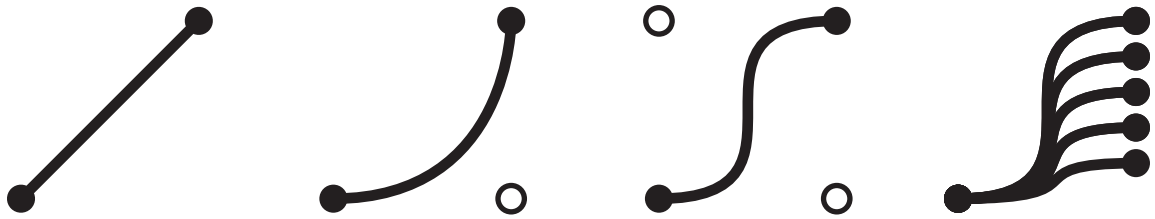


Figure A2: Depiction of building up to a Bézier curve-based continuum. From left to right: (1) A line defined by 2 end points (black circles). (2): A Bézier curve between two end points (black) with curvature defined by the placement of one control point (white). (3): A Bézier curve between two end points with curvature defined by two control points. (4): A continuum of Bézier curves derived by scaling the pitch excursion of the curve.

With regard to the domed onglide, the curved trajectory is implemented using Bézier curves as made available in the `bezier` R package (Olsen, 2018). Bézier curves are a class of Bernstein polynomials[1] with the useful property that a continuous curve between two *endpoints* can be defined by a finite set of *control points* in 2-dimensional space, where the number and position of the

control points affect the curvature. Figure A2 shows a step-by-step schematic of how a continuum for curved L+H* onglides is built up from a simple line through the inclusion of control points.

A.4 Implementation of the Free-Text Response Task

This subsection addresses details of the structure and instructions of the free-text portion of the task. Logistically, forcing participants to write a free-text response immediately upon responding Other would likely discourage them from using this response option. To circumvent this, the experiment back end keeps track of two sets of items as the experiment progresses: (1) the set of continuum steps that the participant responds Other to during the experiment and (2) a hard-coded set of specific continuum steps that *a priori* may yield insightful responses.[2] After completing the three-alternative forced choice portion of the task, the participant then starts the free-text response portion with the following instructions:

> Next, you will listen once more to a few audio clips where either you or other people responded with "other". There will be no more than 20 clips, and you will not need to count aloud. We will ask you to briefly expand on what you think the speaker is trying to convey. This might be distinct from "asking" or "telling," or might be one of these options plus some important nuance. These instructions will be repeated at the top of each page.

The main thing to point out in these instructions is that for the set of trials that the participant responded OTHER to, they are instructed that **they** responded OTHER to that trial. For the hard-coded set interested in but the participant did not respond OTHER to, they are instructed to think of what another **other people** might think. When a participant **does** respond OTHER to one of

[1]The details of the mathematical basis for Bézier curves is beyond the scope of this work and I make no commitment whatsoever to the premise that the implementation of F0 trajectories in natural language are best described in terms of such polynomials. The point here is that Bézier curves provide an easy and flexible way to define curves that is abstracted away from working directly with its mathematical basis (cf. manipulating sigmoidal functions as in Barnes et al., 2021). The reader is directed to seminal work from Bernstein, 1912 and de Faget de Casteljau (1986), modern reviews such as Chapter 7 of Phillips (2006), and public resources like the Wikipedia page for Bézier curves for additional information.

[2]The latter set is particularly important when (1) a participant never responds Other to any trial or (2) participants vary substantially in which continuum steps they respond Other to. Using a **hypothetical** experiment size, if there are 20 steps and 20 participants and each person responds Other to only one step, but each person does so to a different step, then in total we would only get 1 response for each continuum step, which is not very valuable.

the hard-coded continuum steps, that step is moved to the former set of trials. In other words, participants do not give two text responses to the same continuum step and, when possible, the instructions that **they**, rather than other people, responded OTHER is prioritized.[3] Audio clips are arranged as a grid of cards on the screen, with each card containing a play button to play the audio clip and a small textbox below it to respond with. Participants are allowed to freely listen to each audio clip as many times as desired.

[3]In the case that participants respond OTHER multiple times to the same continuum step, but instantiated by different sentences, the most recent step to receive an OTHER response is replayed during the free-text portion. Similarly, for hard-coded steps that did not receive OTHER interpretations, the most recently presented sentence for that continuum step is replayed.

Appendix B

APPENDIX FOR CHAPTER 3 (COMPOSITE MEASURES)

B.1 Formulas

A listing of the model formulas, provided as the right hand side (i.e., the independent variables) of an R formula,[1]. is provided in Table B1. The left hand side (i.e., the dependent variable) is always the binary choice between Telling (=1) or Asking (=0). All models are Bayesian logistic mixed effects regression models fit with the `brms` R package (Bürkner, 2021; Gabry et al., 2024; R Core Team, 2024). The `Shape` predictor (shortened to Sh) is a 2-level categorical variable that is scaled sum coded ($\pm.5$) using the `contrastable` R package (Sostarics, 2024). Other acronyms for variables include Accentual Pitch (`AP`), Ending Pitch (`EP`), Alignment (`AL`),[2] and participant (`ppt`); `item` denotes the 5 sentences used (e.g., *Molly's from Branning*).

B.1.1 Regarding quadratic relationships

The Composite+EP models can be rewritten to contain quadratic relationships that are not immediately obvious. This section provides short derivations for the excursion and TCoG models. For a more extensive discussion of how these quadratic relationships can occur, especially in the context of Simpson's paradox and moderation, see Kock & Gaskins (2016).

A quadratic relationship is easiest to show in the excursion model. In a model using excursion (x_{Exc}), ending pitch (x_{EP}), and their interaction, the interaction term is the excursion times the ending pitch ($x_{Exc}x_{EP}$). Excursion is defined as ending pitch minus accentual pitch ($x_{Exc} = x_{EP} - x_{AP}$). The interaction term can then be expanded to $(x_{EP} - x_{AP})x_{EP}$, simplifying to $\boldsymbol{x_{EP}^2} - x_{AP}x_{EP}$. This algebraic derivation can be taken further for the fixed excursion effect, ultimately resulting in a model structure including predictors of ending pitch, accentual pitch, the

[1]For the unfamiliar reader, for two predictors A and B, `A+B` denotes fixed effects of A and B; `A:B` indicates an interaction term between A and B; `A*B` is a shorthand for `A + B + A:B`; for a term `C` denoting a clustering variable (such as participant), `(1|C)` denotes varying intercepts by cluster C; `(1+A|C)` denotes varying slopes of A by C **and** varying intercepts by C. `I(X^2)` denotes a quadratic predictor of `X`.

[2]Alignment is represented as % stressed syllable duration away from the stressed syllable right boundary

Name	Formula
Scaling	`AP*EP + (1+AP*EP|ppt) + (1|item)`
Scaling+AL	`AP*EP + AL+EP:AL + (1+AP*EP+AL+EP:AL|ppt) + (1|item)`
Scaling+AP^2	`AP*EP + I(AP^2) + (1+AP*EP+I(AP^2)|ppt) + (1|item)`
Scaling+EP^2	`AP*EP + I(EP^2) + (1+AP*EP+I(EP^2)|ppt) + (1|item)`
Excursion	`Excursion + (1+Excursion|ppt) + (1|item)`
Excursion+Sh	`Excursion*Sh + (1+Excursion*Sh|ppt) + (1|item)`
Excursion+AP	`Excursion*AP + (1+Excursion*AP|ppt) + (1|item)`
Excursion+EP	`Excursion*EP + (1+Excursion*EP|ppt) + (1|item)`
Slope	`Slope + (1+Slope|ppt) + (1|item)`
Slope+Sh	`Slope*Sh + (1+Slope*Sh|ppt) + (1|item)`
Slope+AP	`Slope*AP + (1+Slope*AP|ppt) + (1|item)`
Slope+EP	`Slope*EP + (1+Slope*EP|ppt) + (1|item)`
TCoG	`TCoG + (1+TCoG|ppt) + (1|item)`
TCoG+Sh	`TCoG*Sh + (1+TCoG*Sh|ppt) + (1|item)`
TCoG+AP	`TCoG*AP + (1+TCoG*AP|ppt) + (1|item)`
TCoG+EP	`TCoG*EP + (1+TCoG*EP|ppt) + (1|item)`

Table B1: R formulas for models used in Chapter 3

interaction between ending and accentual pitch, **and ending pitch squared**. Crucially, a model using these first three terms is what defines the scaling model; the only difference between the two is whether some of the coefficients are constrained to be equal to one another.[3] The slope model would have the same derivation as the excursion model, but the magnitude would be amended by a scaling factor related to the temporal duration.

Showing a quadratic relationship in the TCoG model requires a few more steps, but the relevant insight is that TCoG, by definition, is a weighted sum where the final F0 sample receives the highest weight. When the weights correspond to time-normalized timestamps (as they do in this work), the final sample's weight is 1.0. The final F0 sample, by definition, is the ending pitch value x_{EP}; thus the final term in the weighted sum is just the ending pitch value ($1 \cdot x_{EP}$). When this is multiplied

[3]For example, the Excursion+EP model would expand to $(\beta_{Exc} + \beta_{EP})x_{EP} - \beta_{Exc}x_{AP} - \beta_{Exc:EP}x_{AP}x_{EP} + \beta_{Exc:EP}x_{EP}^2$. Here, the coefficient for $x_{AP}x_{EP}$ and x_{EP}^2 have the same magnitude but differing signs. Explicitly fitting a model with all of these terms, as with the Scaling+EP model, does not have this constraint and so allows it to be slightly more flexible. That said, the two models are nonetheless comparable in terms of the structure of their predictors. A likelihood ratio test comparing the Excursion+EP model and a separate model explicitly using the expanded form shows that the two models are not different from one another.

	Scaling			Scaling+AL		
Term	Estimate	(SE)	95% CrI	Estimate	(SE)	95% CrI
Intercept	2.12	(0.201)	[1.76, 2.56]	2.15	(0.203)	[1.78, 2.59]
AccentualPitch	−0.04	(0.017)	[−0.08, −0.01]	−0.05	(0.018)	[−0.08, −0.01]
EndingPitch	−0.65	(0.020)	[−0.69, −0.61]	−0.63	(0.020)	[−0.67, −0.59]
:AccentualPitch	0.03	(0.004)	[0.02, 0.03]	0.02	(0.004)	[0.01, 0.02]
Alignment				0.00	(0.036)	[−0.07, 0.07]
:EndingPitch				0.01	(0.012)	[−0.02, 0.03]
Random Effects by Participant						
Intercept	1.16	(0.063)	[1.04, 1.29]	1.19	(0.066)	[1.06, 1.32]
AccentualPitch	0.24	(0.016)	[0.21, 0.27]	0.24	(0.017)	[0.21, 0.28]
EndingPitch	0.32	(0.018)	[0.29, 0.36]	0.30	(0.018)	[0.27, 0.34]
:AccentualPitch	0.05	(0.004)	[0.04, 0.05]	0.04	(0.004)	[0.03, 0.05]
Alignment				0.20	(0.055)	[0.08, 0.30]
:EndingPitch				0.09	(0.012)	[0.07, 0.11]
Random Effects by Sentence						
Intercept	0.43	(0.153)	[0.21, 0.80]	0.43	(0.151)	[0.22, 0.80]

Table B2: Model summary for the Scaling and Scaling+AL models.

by x_{EP} in the interaction term, we again obtain a quadratic term of x_{EP}^2. Similarly, a term can be pulled out from the earlier in the sum corresponding to the accentual pitch target which will additionally yield an $x_{AP}x_{EP}$ term via the interaction term, although the weight of this term will be lower than the weight given to the final term (i.e., the weight given to ending pitch). Note that this is setting aside the $\sum t_i$ from the denominator, which is just a constant scaling factor.

B.2 Model Summary Tables

This section presents model summaries for each of the models reported in this work. Due to space constraints, two models are reported per table. Note that the random effect estimates describe a standard deviation parameter, which is strictly positive. Interaction terms are shown nested beneath the first term in the interaction. Model formulas are given in B.1. Table B10 additionally lists all of the model metrics reported in this work. For the +Shape models, *Rising-Falling* refers to the categorical difference between the rising and the falling contours (collectively).

	Scaling+AP^2			**Scaling+EP^2**		
Term	Estimate	(SE)	95% CrI	Estimate	(SE)	95% CrI
Intercept	1.97	(0.201)	[1.60, 2.40]	2.42	(0.228)	[2.02, 2.92]
AccentualPitch	−0.04	(0.017)	[−0.07, −0.01]	−0.01	(0.017)	[−0.04, 0.02]
EndingPitch	−0.66	(0.020)	[−0.70, −0.62]	−0.66	(0.024)	[−0.71, −0.61]
:AccentualPitch	0.02	(0.004)	[0.02, 0.03]	0.03	(0.004)	[0.02, 0.04]
AP^2	0.02	(0.003)	[0.01, 0.03]			
EP^2				−0.02	(0.003)	[−0.02, −0.01]
Random Effects by Participant						
Intercept	1.12	(0.064)	[1.00, 1.26]	1.10	(0.062)	[0.99, 1.23]
AccentualPitch	0.23	(0.016)	[0.20, 0.26]	0.23	(0.015)	[0.20, 0.26]
EndingPitch	0.33	(0.018)	[0.29, 0.36]	0.35	(0.021)	[0.31, 0.39]
:AccentualPitch	0.05	(0.004)	[0.04, 0.05]	0.05	(0.004)	[0.04, 0.05]
AP^2	0.03	(0.004)	[0.02, 0.04]			
EP^2				0.03	(0.002)	[0.03, 0.04]
Random Effects by Sentence						
Intercept	0.43	(0.152)	[0.22, 0.80]	0.47	(0.163)	[0.24, 0.86]

Table B3: Model summary for the Scaling+AP^2 and Scaling+EP^2 models.

	Excursion			**Excursion+Shape**		
Term	Estimate	(SE)	95% CrI	Estimate	(SE)	95% CrI
Intercept	1.03	(0.150)	[0.78, 1.37]	1.74	(0.164)	[1.45, 2.09]
Excursion	−0.39	(0.013)	[−0.41, −0.36]	−0.27	(0.017)	[−0.30, −0.24]
Rising-Falling				−1.84	(0.126)	[−2.09, −1.59]
:Excursion				−0.22	(0.033)	[−0.29, −0.16]
Random Effects by Participant						
Intercept	0.93	(0.041)	[0.85, 1.01]	1.15	(0.062)	[1.03, 1.28]
Excursion	0.22	(0.011)	[0.20, 0.24]	0.19	(0.014)	[0.17, 0.22]
Rising-Falling				1.55	(0.104)	[1.35, 1.76]
:Excursion				0.41	(0.027)	[0.36, 0.47]
Random Effects by Sentence						
Intercept	0.36	(0.150)	[0.17, 0.74]	0.38	(0.154)	[0.18, 0.77]

Table B4: Model summary for the Excursion and Excursion+Shape models.

	Excursion+AP			Excursion+EP		
Term	Estimate	(SE)	95% CrI	Estimate	(SE)	95% CrI
Intercept	2.03	(0.145)	[1.75, 2.32]	2.45	(0.176)	[2.14, 2.83]
Excursion	−0.63	(0.020)	[−0.67, −0.59]	−0.01	(0.017)	[−0.04, 0.02]
AccentualPitch	−0.60	(0.019)	[−0.64, −0.57]			
:Excursion	0.00	(0.003)	[0.00, 0.01]			
EndingPitch				−0.63	(0.020)	[−0.67, −0.59]
:Excursion				−0.02	(0.003)	[−0.03, −0.02]
Random Effects by Participant						
Intercept	1.06	(0.058)	[0.95, 1.18]	1.05	(0.058)	[0.94, 1.17]
Excursion	0.32	(0.017)	[0.28, 0.35]	0.24	(0.014)	[0.22, 0.27]
AccentualPitch	0.26	(0.016)	[0.23, 0.29]			
:Excursion	0.04	(0.003)	[0.03, 0.04]			
EndingPitch				0.30	(0.018)	[0.27, 0.34]
:Excursion				0.03	(0.002)	[0.03, 0.04]
Random Effects by Sentence						
Intercept	0.39	(0.141)	[0.20, 0.75]	0.47	(0.171)	[0.23, 0.89]

Table B5: Model summary for the Excursion+AP and Excursion+EP models.

	Slope			Slope+Shape		
Term	Estimate	(SE)	95% CrI	Estimate	(SE)	95% CrI
Intercept	0.31	(0.128)	[0.04, 0.54]	1.42	(0.144)	[1.15, 1.72]
Slope	−5.71	(0.238)	[−6.18, −5.24]	−4.22	(0.211)	[−4.63, −3.81]
Rising-Falling				−1.79	(0.107)	[−2.01, −1.59]
:Slope				−5.68	(0.347)	[−6.35, −4.99]
Random Effects by Participant						
Intercept	0.91	(0.040)	[0.84, 1.00]	1.01	(0.052)	[0.91, 1.12]
Slope	4.62	(0.194)	[4.26, 5.01]	3.05	(0.156)	[2.75, 3.36]
Rising-Falling				1.17	(0.087)	[1.00, 1.34]
:Slope				5.90	(0.263)	[5.40, 6.43]
Random Effects by Sentence						
Intercept	0.31	(0.131)	[0.15, 0.65]	0.34	(0.137)	[0.17, 0.69]

Table B6: Model summary for the Slope and Slope+Shape models.

	Slope+AP			Slope+EP		
Term	Estimate	(SE)	95% CrI	Estimate	(SE)	95% CrI
Intercept	1.05	(0.165)	[0.69, 1.34]	2.18	(0.172)	[1.88, 2.55]
Slope	−7.97	(0.381)	[−8.71, −7.22]	−0.72	(0.217)	[−1.15, −0.30]
AccentualPitch	−0.45	(0.019)	[−0.48, −0.41]			
:Slope	0.62	(0.090)	[0.45, 0.80]			
EndingPitch				−0.60	(0.020)	[−0.64, −0.56]
:Slope				−0.27	(0.036)	[−0.34, −0.20]
Random Effects by Participant						
Intercept	1.01	(0.053)	[0.91, 1.11]	1.20	(0.059)	[1.09, 1.32]
Slope	6.79	(0.275)	[6.27, 7.35]	3.63	(0.177)	[3.29, 3.99]
AccentualPitch	0.26	(0.014)	[0.23, 0.29]			
:Slope	1.24	(0.076)	[1.09, 1.39]			
EndingPitch				0.31	(0.016)	[0.28, 0.34]
:Slope				0.48	(0.030)	[0.42, 0.54]
Random Effects by Sentence						
Intercept	0.41	(0.160)	[0.19, 0.82]	0.45	(0.166)	[0.23, 0.86]

Table B7: Model summary for the Slope+AP and Slope+EP models.

	TCoG			TCoG+Shape		
Term	Estimate	(SE)	95% CrI	Estimate	(SE)	95% CrI
Intercept	1.39	(0.154)	[1.12, 1.73]	1.91	(0.160)	[1.62, 2.25]
TCoG	−0.76	(0.021)	[−0.80, −0.72]	−0.54	(0.025)	[−0.59, −0.49]
Rising-Falling				−2.32	(0.100)	[−2.52, −2.13]
:TCoG				−0.40	(0.034)	[−0.47, −0.33]
Random Effects by Participant						
Intercept	0.96	(0.046)	[0.87, 1.05]	0.96	(0.050)	[0.87, 1.06]
TCoG	0.35	(0.016)	[0.32, 0.38]	0.40	(0.021)	[0.36, 0.44]
Rising-Falling				1.41	(0.086)	[1.25, 1.59]
:TCoG				0.40	(0.031)	[0.34, 0.46]
Random Effects by Sentence						
Intercept	0.35	(0.152)	[0.16, 0.74]	0.40	(0.155)	[0.19, 0.79]

Table B8: Model summary for the TCoG and TCoG+Shape models.

	TCoG+AP			TCoG+EP		
Term	Estimate	(SE)	95% CrI	Estimate	(SE)	95% CrI
Intercept	1.65	(0.142)	[1.36, 1.93]	2.24	(0.158)	[1.96, 2.57]
TCoG	−1.21	(0.039)	[−1.29, −1.13]	0.12	(0.040)	[0.04, 0.20]
AccentualPitch	0.31	(0.029)	[0.25, 0.36]			
:TCoG	0.05	(0.006)	[0.04, 0.06]			
EndingPitch				−0.67	(0.029)	[−0.73, −0.61]
:TCoG				−0.02	(0.004)	[−0.03, −0.01]
Random Effects by Participant						
Intercept	1.15	(0.060)	[1.04, 1.27]	1.13	(0.057)	[1.02, 1.24]
TCoG	0.63	(0.034)	[0.57, 0.70]	0.63	(0.033)	[0.57, 0.70]
AccentualPitch	0.48	(0.023)	[0.44, 0.53]			
:TCoG	0.06	(0.006)	[0.05, 0.07]			
EndingPitch				0.47	(0.026)	[0.42, 0.52]
:TCoG				0.05	(0.003)	[0.04, 0.06]
Random Effects by Sentence						
Intercept	0.35	(0.134)	[0.18, 0.69]	0.42	(0.151)	[0.21, 0.79]

Table B9: Model summary for the TCoG+AP and TCoG+EP models.

Model	**AUC**	**95% CI**	z	p	**ELPD**	**(SE)**	**Diff**	**(SE)**
Scaling+EP^2	0.938	[0.936, 0.940]			−13 131	(119.5)		
Excursion+EP	0.938	[0.935, 0.940]	−0.87	0.386	−13 156	(119.8)	−26	(9.1)
Slope+EP	0.936	[0.934, 0.939]	−5.48	<0.001	−13 302	(121.1)	−171	(32.2)
TCoG+EP	0.936	[0.934, 0.938]	−6.53	<0.001	−13 326	(121.9)	−196	(32.1)
Scaling+AL	0.932	[0.930, 0.935]	−12.96	<0.001	−13 755	(123.5)	−624	(46.9)
Scaling+AP^2	0.932	[0.929, 0.934]	−15.44	<0.001	−13 820	(123.1)	−689	(44.1)
Scaling	0.931	[0.929, 0.934]	−17.25	<0.001	−13 858	(122.6)	−727	(42.8)
Excursion+AP	0.931	[0.929, 0.934]	−15.81	<0.001	−13 868	(124.0)	−737	(47.0)
TCoG+Shape	0.924	[0.921, 0.926]	−20.29	<0.001	−14 361	(116.0)	−1230	(66.8)
TCoG+AP	0.926	[0.923, 0.928]	−24.37	<0.001	−14 384	(122.6)	−1253	(51.9)
Slope+AP	0.927	[0.924, 0.929]	−20.39	<0.001	−14 689	(133.5)	−1558	(74.6)
Excursion+Shape	0.899	[0.897, 0.902]	−39.70	<0.001	−16 311	(113.1)	−3180	(78.1)
Slope+Shape	0.899	[0.896, 0.902]	−39.24	<0.001	−16 386	(109.9)	−3255	(79.6)
TCoG	0.894	[0.891, 0.897]	−42.70	<0.001	−16 740	(114.1)	−3609	(81.2)
Excursion	0.889	[0.886, 0.892]	−45.60	<0.001	−17 311	(118.3)	−4180	(83.0)
Slope	0.884	[0.881, 0.888]	−46.82	<0.001	−18 132	(136.9)	−5001	(104.5)

Table B10: Listing of performance metrics for each model. Models are listed in order of decreasing performance. ROC comparisons (z and corresponding p values) are made using the DeLong method in the `pROC` R package (Robin et al., 2011). ROC comparisons and ELPD differences are shown relative to the Scaling+EP^2 model.

B.3 Supplementary Figures

Figure B1 shows the predictions for all 16 models without including variation from the random effects. For ease of comparison, Figure B2 shows the predictions for all 16 models when random effects are included (these figures were presented individually in the main text).

The main text presented the area under the receiver operating characteristic (ROC) curve. The actual ROC curves are displayed in Figure B3. Although the individual ROC curves from each fold from the 5-fold cross validation procedure are plotted (as light lines), they do not differ substantially from one another. The darker lines overlaid on top are the ROC curves averaged[4] across the five folds.

[4]Averaging is done vertically, which maintains the average AUC across the five folds in the averaged curve. For a discussion of other ways to average ROC curves, see Hogan & Adams (2023).

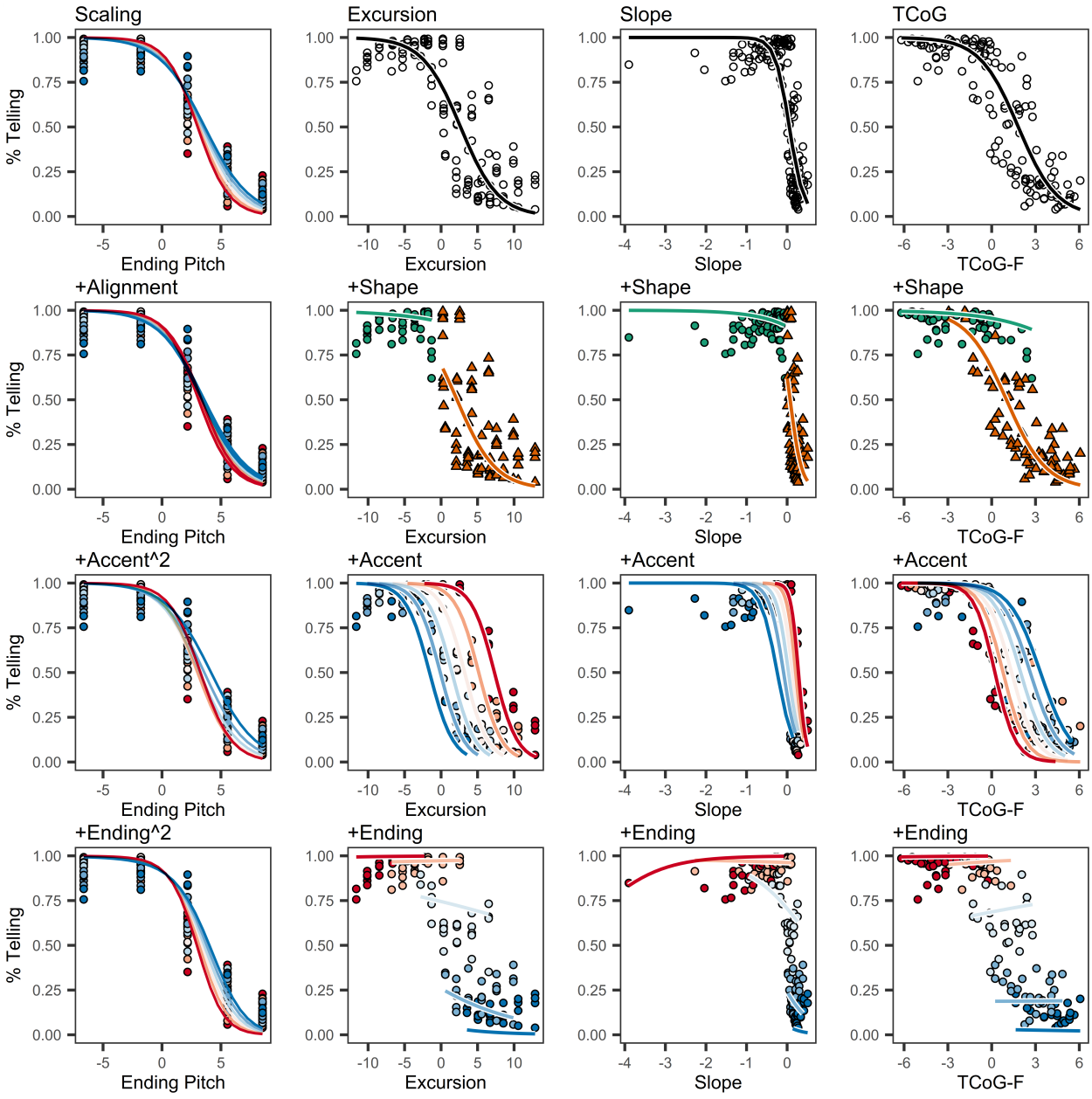


Figure B1: Model predictions versus empirical data without including variation from the random effects. One point equals the average proportion of Telling responses for one contour from each experiment. For models including rising/falling shape as a predictor, rises are shown with orange triangles while falls are shown with green circles. For the left column and third row, points are colored by accentual pitch step (low=red, high=blue). For the three composite plots in the bottom row, points are colored by ending pitch step (low=red, high=blue).

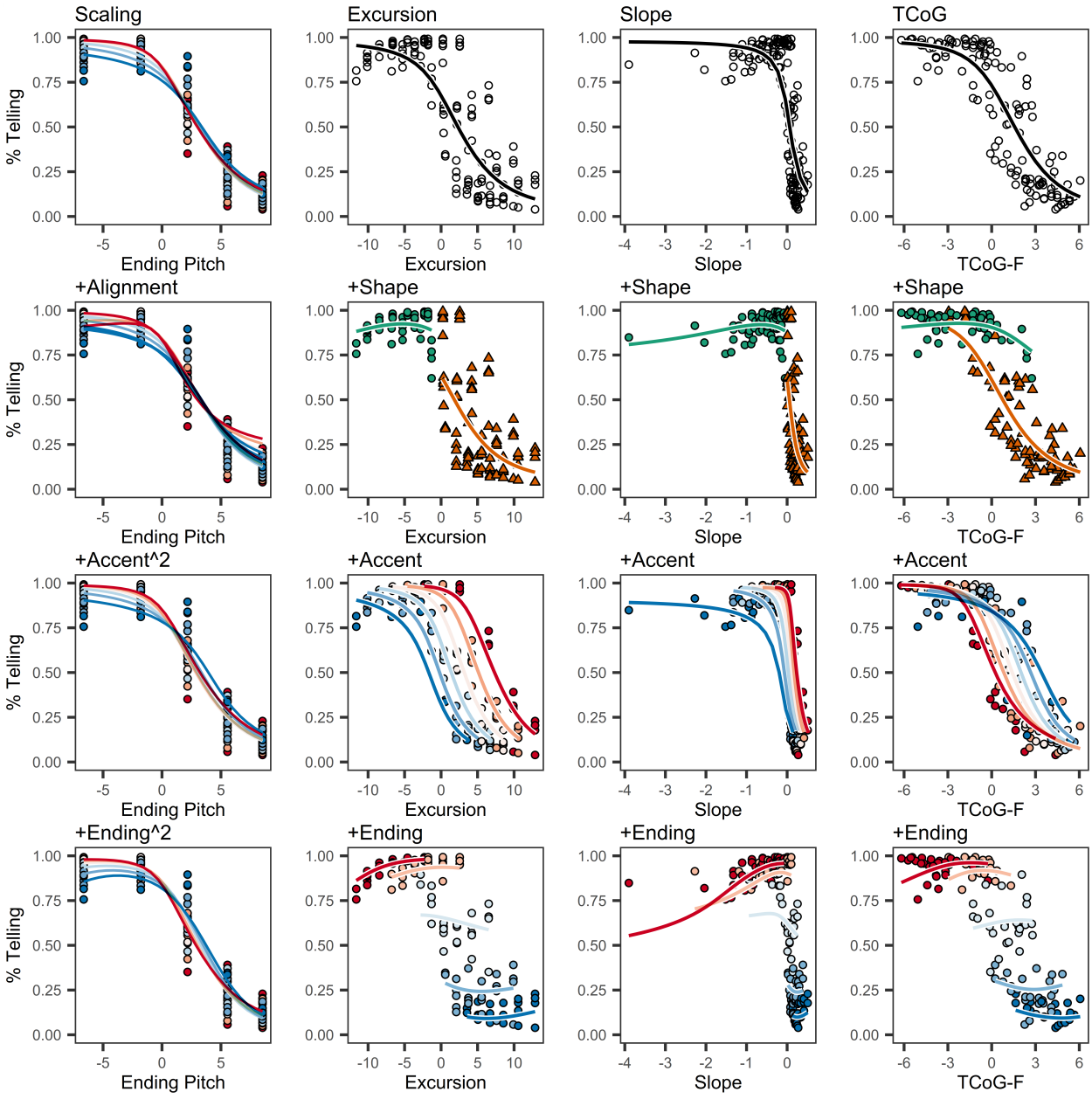
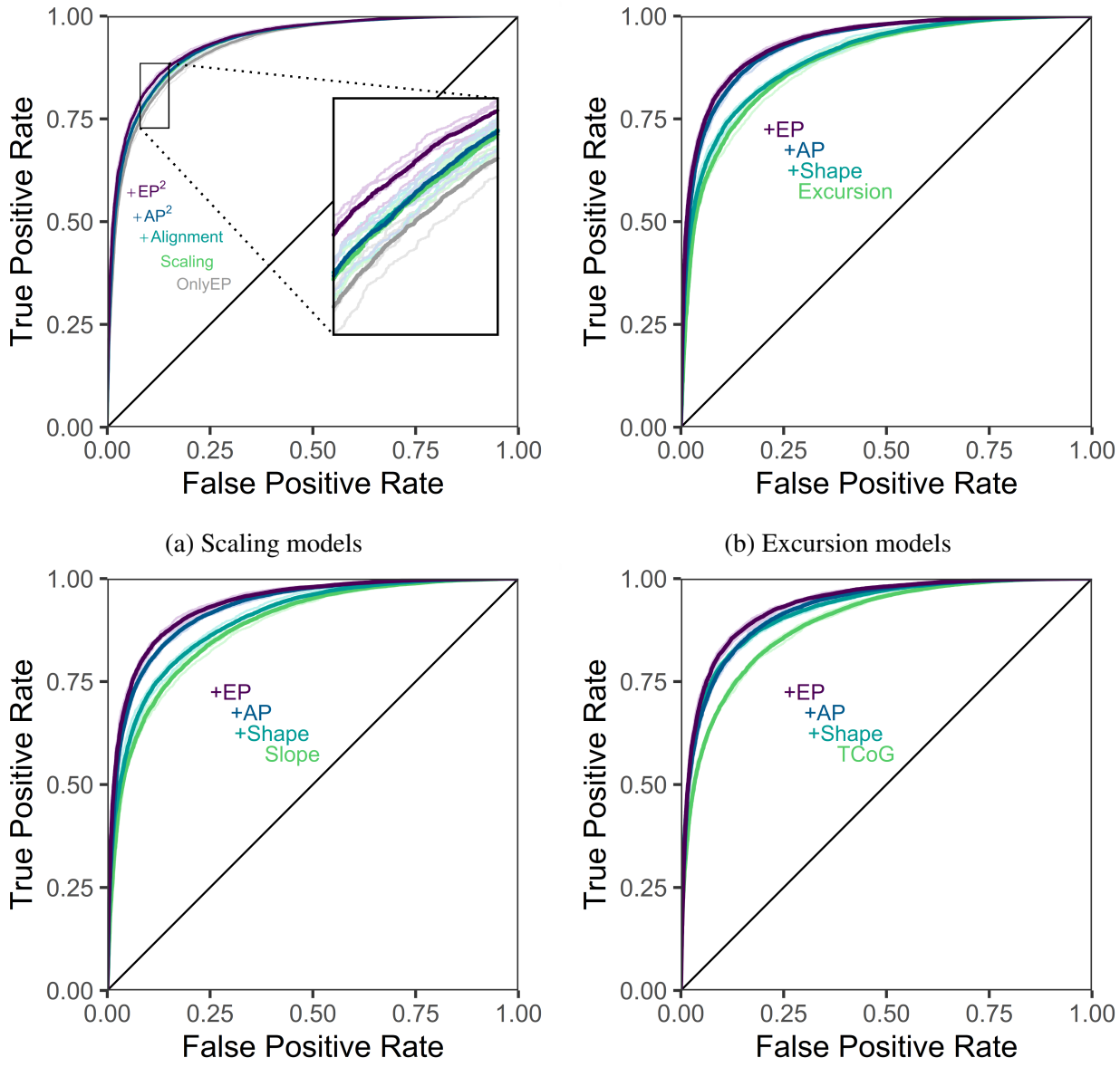


Figure B2: Model predictions versus empirical data with random effects included. One point equals the average proportion of Telling responses for one contour from each experiment. For models including rising/falling shape as a predictor, rises are shown with orange triangles while falls are shown with green circles. For the left column and third row, points are colored by accentual pitch step (low=red, high=blue). For the three composite plots in the bottom row, points are colored by ending pitch step (low = red, high = blue).



(a) Scaling models

(b) Excursion models

(c) Slope models

(d) TCoG models

Figure B3: ROC curves for the each statistical model. Generally, better performance is indicated by curves that bend closer to the upper-left corner.

Appendix C

APPENDIX FOR CHAPTER 4 (RISE-FALL-RISE)

Note that this appendix contains a large number of tables and figures that take up the entire page. To assist with cross references in paragraphs–where the referenced figure or table may be several pages away–the page number that the figure or table appears on will also be included in the text.

C.1 Dialogue Materials

This section lists the materials for each dialogue used in the experiments in these works. The key is as follows: ***Target word(s)*: Question? Answer(s). Prompt(s).** For the critical items, the LowerTarget condition and HigherTarget condition answers and prompts are separated by slashes (i.e., LowerTarget/HigherTarget). Items are not ordered in any particular order.

C.1.1 Critical items

All answers are recoded with all 6 items. Targets for the LowerTarget and HigherTarget conditions are shown as the relevant scale, i.e., <*LowerTarget*, *HigherTarget*>.

1. <*understandable*, *articulate*>: Are the classes with the old professor worth taking? He's articulate / He's understandable.
   Would you conclude that the professor is not merely understandable? / Would you conclude that the professor is not articulate?
2. <*adequate*, *good*>: Are you done editing your final paper yet? The writing is good / The writing is good / The writing is adequate.
   Would you conclude that his writing is not merely adequate? / Would you conclude that their writing is not merely adequate? / Would you conclude that his writing is not good?
3. <*wary*, *scared*>: Can your son cross the street all by himself? Jimmy still gets scared / Jimmy still gets wary / Jimmy still gets wary.
   Would you conclude that Jimmy does not merely get wary? / Would you conclude that Jimmy does not get scared? / Would you conclude that Jimmy is not scared?
4. <*happy*, *delighted*>: Did Jane like the surprise party we threw her? The gesture made her delighted / The gesture made her happy.
   Would you conclude that Jane was not merely happy? / Would you conclude that Jane was not delighted?

5. <*cold*, *freezing*>: Did someone leave a window open in the office overnight? The office feels freezing / The office feels cold / The office feels cold.
   Would you conclude that the office does not merely feel cold? / Would you conclude that the office does not feel freezing? / Would you conclude that the office is not freezing

6. <*fat*, *obese*>: Did that actor put on weight for the role? In the trailer he looks obese / In the trailer he looks fat / In the trailer he looks fat.
   Would you conclude that the actor did not merely look fat? / Would you conclude that the actor did not look obese? / Would you conclude that the actor was not obese?

7. <*large*, *gigantic*>: Did the university add space for the grad student offices? The addition is gigantic / The addition is large.
   Would you conclude that the addition is not merely large? / Would you conclude that the addition is not gigantic?

8. <*difficult*, *impossible*>: Did you come up with anything for the last problem on the exam? That one was impossible / That one was difficult.
   Would you conclude that the problem was not merely difficult? / Would you conclude that the problem was not impossible?

9. <*hard*, *unsolvable*>: Did you finish this week's chemistry homework? The last section was unsolvable / The last section was hard.
   Would you conclude that the last section was not merely hard? / Would you conclude that the last section was not unsolvable?

10. <*warm*, *hot*>: Did you leave the pool heater running again? The pool water feels hot / The pool water feels warm.
    Would you conclude that the water does not merely feel warm? / Would you conclude that the water does not feel hot?

11. <*pretty*, *beautiful*>: Did you see that sunset yesterday? The sunset was beautiful / The sunset was pretty.
    Would you conclude that the sunset was not merely pretty? / Would you conclude that the sunset was not beautiful?

12. <*big*, *enormous*>: Did you see the baby elephant at the zoo? That animal was enormous / That animal was big.
    Would you conclude that the elephant was not merely big? / Would you conclude that the elephant was not enormous?

13. <*annoyed*, *angry*>: Did you see the deans in the lobby this morning? Dean Johnson seemed angry / Dean Johnson seemed annoyed.
    Would you conclude that the dean is not merely annoyed? / Would you conclude that the dean is not angry?

14. <*unkind*, *nasty*>: Did you see the teaching evaluations for the class? The students were nasty / The students were unkind.

Would you conclude that the students were not merely unkind? / Would you conclude that the students were not nasty?

15. <*special*, *unique*>: Did you try the cake from the new bakery? The flavors were unique / The flavors were special.
Would you conclude that the flavors were not merely special? / Would you conclude that the flavors were not unique?

16. <*nice*, *great*>: Did your graduation ceremony really still happen even with the thunderstorm? The event was still great / The event was still nice.
Would you conclude that the event was not merely nice? / Would you conclude that the event was not great?

17. <*polite*, *friendly*>: Did your meeting with the new professor go well? The professor was friendly / The professor was polite.
Would you conclude that the professor was not merely polite? / Would you conclude that the professor was not friendly?

18. <*scared*, *petrified*>: Did your niece go on the big roller coasters at six flags? Little Alice was petrified / Little Alice was scared.
Would you conclude that Alice was not merely scared? / Would you conclude that Alice was not petrified?

19. <*happy*, *ecstatic*>: Did your wife react well when the doctor said you were having triplets? Marilyn was ecstatic / Marilyn was happy.
Would you conclude that she was not merely happy? / Would you conclude that she was not ecstatic?

20. <*attractive*, *stunning*>: Didn't your ex-girlfriend win a beauty pageant? Natalie's stunning / Natalie's attractive.
Would you conclude that his ex-girlfriend is not merely attractive? / Would you conclude that his ex-girlfriend is not stunning?

21. <*soft*, *mushy*>: Do the avocados on the counter need to be thrown out? The avocados feel mushy / The avocados feel soft.
Would you conclude that the avocados do not merely feel soft? / Would you conclude that the avocados do not feel mushy?

22. <*snug*, *tight*>: Do those hand-me-down clothes fit the kids? The blue sweater is tight / The blue sweater is snug.
Would you conclude that the sweater is not merely snug? / Would you conclude that the sweater is not tight?

23. <*gray*, *black*>: Do you have a suit you could wear to the funeral? My business suit is black / My business suit is gray.
Would you conclude that his suit is not merely gray? / Would you conclude that his suit is not black?

24. <*thin*, *invisible*>: Do you have any scratches on your phone? There's one on the front that's invisible / There's one on the front that's thin.
Would you conclude that the scratch is not merely thin? / Would you conclude that the scratch is not invisible?

25. <*unsettling*, *horrific*>: Do you like zombie movies? They're horrific / They're unsettling.
Would you conclude that the movies are not merely unsettling? / Would you conclude that the movies are not horrific?

26. <*quiet*, *silent*>: Do you think I could record some lines for my theatre class in the office? The office is silent / The office is quiet.
Would you conclude that the office is not merely quiet? / Would you conclude that the office is not silent?

27. <*smart*, *brilliant*>: Do you think the new research assistant could help out with that difficult analysis? That assistant is brilliant / That assistant is smart.
Would you conclude that the assistant is not merely smart? / Would you conclude that the assistant is not brilliant?

28. <*likely*, *certain*>: Do you think the union will secure a raise for our salary? A raise is certain / A raise is likely.
Would you conclude that a raise is not merely likely? / Would you conclude that a raise is not certain?

29. <*thick*, *impenetrable*>: Do you think this coat will survive the chicago winter? The material feels impenetrable / The material feels thick.
Would you conclude that the material does not merely feel thick? / Would you conclude that the material does not feel impenetrable?

30. <*busy*, *full*>: Do you think we'll be able to get a table at this restaurant without a reservation? This place seems full / This place seems busy.
Would you conclude that the restaurant does not seem merely busy? / Would you conclude that the restaurant does not seem full?

31. <*hungry*, *starving*>: Do you want to go to that new sushi place? I'm feeling starving / I'm feeling hungry.
Would you conclude that he does not merely feel hungry? / Would you conclude that he does not feel starving?

32. <*chubby*, *fat*>: Does the baby still fit those clothes? The baby is fat / The baby is chubby.
Would you conclude that the baby is not merely chubby? / Would you conclude that the baby is not fat?

33. <*mediocre*, *bad*>: Does the intern understand what he's supposed to do? His work is bad / His work is mediocre.
Would you conclude that the intern's work is not merely mediocre? / Would you conclude that the intern's work is not bad?

34. <*cool*, *cold*>: Has the air conditioning been on all day? The living room feels cold / The living room feels cool.
Would you conclude that the living room does not merely feel cool? / Would you conclude that the living room does not feel cold?

35. <*rough*, *unfriendly*>: Has the new rescue dog been getting along with your other dogs? He'll play but he's unfriendly / He'll play but he's rough.
Would you conclude that the dog is not merely rough? / Would you conclude that the dog is not unfriendly?

36. <*scarce*, *unavailable*>: Has the supply of baby formula gone back to normal? Baby formula is unavailable / Baby formula is scarce.
Would you conclude that baby formula is not merely scarce? / Would you conclude that baby formula is not unavailable?

37. <*old*, *ancient*>: Has your family always lived in this house? Our humble abode is ancient / Our humble abode is old.
Would you conclude that the home is not merely old? / Would you conclude that the home is not ancient?

38. <*unhappy*, *miserable*>: Have you bounced back after breaking up with your girlfriend? I am still miserable / I am still unhappy.
Would you conclude that he is not merely unhappy? / Would you conclude that he is not miserable?

39. <*ugly*, *hideous*>: Have you seen the design for the new stadium? The design is hideous / The design is ugly.
Would you conclude that the design is not merely ugly? / Would you conclude that the design is not hideous?

40. <*calm*, *meditative*>: Have you tried out the new yoga class? The atmosphere was meditative / The atmosphere was calm.
Would you conclude that the atmosphere was not merely calm? / Would you conclude that the atmosphere was not meditative?

41. <*low*, *depleted*>: Hey you got the new iPhone too, right? Does your battery last the whole day? The battery is always depleted / The battery is always low.
Would you conclude that the battery is not merely low? / Would you conclude that the battery is not depleted?

42. <*red*, *scarlet*>: I accidentally washed my shirt on the wrong setting, do you think the colors washed out a bit? The t shirt is still scarlet / The t shirt is still red.
Would you conclude that the the t shirt is not merely red? / Would you conclude that the t shirt is not scarlet?

43. <*tough*, *impossible*>: I haven't gone running since before the pandemic, do you think I could do a half marathon? That distance would be impossible / That distance would be tough.

Would you conclude that that distance is not merely tough? / Would you conclude that that distance is not impossible?

44. <*strenuous*, *exhausting*>: I heard today was a record high with over 80Would you conclude that the hike was not merely strenuous? / Would you conclude that the hike was not exhausting?

45. <*dirty*, *filthy*>: I liked that carpet we saw at the garage sale, should we buy it? That carpet looked filthy / That carpet looked dirty.
Would you conclude that the carpet did not merely look dirty? / Would you conclude that the carpet did not look filthy?

46. <*dark*, *black*>: I missed chemistry yesterday, do you remember what happens when you add a drop of iodine to the solution? The solution turns black / The solution turns dark.
Would you conclude that the solution does not merely turn dark? / Would you conclude that the solution does not turn black?

47. <*intelligent*, *brilliant*>: I missed the prospective students weekend, did you meet any of the new students? They were brilliant / They were intelligent.
Would you conclude that the students were not merely intelligent? / Would you conclude that the students were not brilliant?

48. <*palatable*, *delicious*>: I was gonna grab something to eat at happy hour, would you recommend the food at the bar? The food is delicious / The food is palatable.
Would you conclude that the food is not merely palatable? / Would you conclude that the food is not delicious?

49. <*honest*, *blunt*>: I'm nervous about my first performance evaluation, have you had yours yet? My bosses were blunt / My bosses were honest.
Would you conclude that the bosses were not merely honest? / Would you conclude that the bosses were not blunt?

50. <*comfortable*, *luxurious*>: I'm traveling to New York for a conference, have you stayed at the Hilton? Their beds are luxurious / Their beds are comfortable.
Would you conclude that the Hilton beds are not merely comfortable? / Would you conclude that the Hilton beds are not luxurious?

51. <*hot*, *boiling*>: Is the kettle for the tea finished yet? The water is boiling / The water is hot.
Would you conclude that the water is not merely hot? / Would you conclude that the water is not boiling?

52. <*cute*, *adorable*>: Jane's dog just had puppies, did she send you the pictures of the pugs? Those pug puppies are adorable / Those pug puppies are cute.
Would you conclude that the puppies are not merely cute? / Would you conclude that the puppies are not adorable?

53. <*pretty*, *gorgeous*>: The ceremony was cloudy, but did the photographer do a good job? The pictures were gorgeous / The pictures were pretty.

Would you conclude that the pictures were not merely pretty? / Would you conclude that the pictures were not gorgeous?

54. <*rare*, *extinct*>: The museum had so many taxidermied birds, do you think we'll ever see them in the wild? The larger ones are extinct / The larger ones are rare.
Would you conclude that the larger birds are not merely rare? / Would you conclude that the larger birds are not extinct?

55. <*loud*, *deafening*>: Was it fun being part of the crowd at the superbowl? The crowd's cheering was deafening / The crowd's cheering was loud.
Would you conclude that the crowd's cheering was not merely loud? / Would you conclude that the crowd's cheering was not deafening?

56. <*enjoyable*, *great*>: Was the VIP meet and greet worth the extra price? It was great / It was enjoyable.
Would you conclude that the meet and greet was not merely enjoyable? / Would you conclude that the meet and greet was not great?

57. <*big*, *huge*>: We need someone to host the prospective students party, does your apartment have enough space? My apartment is huge / My apartment is big.
Would you conclude that the apartment is not merely big? / Would you conclude that the apartment is not huge?

58. <*poor*, *destitute*>: Were the families in the community affected by the stock market crash? The families became destitute / The families became poor.
Would you conclude that the families are not merely become poor? / Would you conclude that the families did not become destitute?

59. <*quiet*, *inaudible*>: Were you able to sleep when you lived in the dorm next to the boiler room? The machines were inaudible / The machines were quiet.
Would you conclude that the machines were not merely quiet? / Would you conclude that the machines were not inaudible?

60. <*small*, *tiny*>: Would John be able to drive all of us to the party? His vehicle is tiny / His vehicle is small.
Would you conclude that the car is not merely small? / Would you conclude that the car is not tiny?

61. <*casual*, *sloppy*>: Would you go on a date at that barbecue place? That place is sloppy / That place is casual.
Would you conclude that the place is not merely casual? / Would you conclude that the place is not sloppy?

62. <*funny*, *hilarious*>: Would you recommend that comedy club you went to last week? The improv act was hilarious / The improv act was funny.
Would you conclude that the improv act was not merely funny? / Would you conclude that the improv act was not hilarious?

63. <*satisfactory*, *impeccable*>: Would you recommend those movers you hired last fall? Their service was impeccable / Their service was satisfactory.
Would you conclude that their service was not merely satisfactory? / Would you conclude that their service was not impeccable?

64. <*good*, *excellent*>: You have T-Mobile, right? I'm thinking of switching, do you get service downtown? The signal there is excellent / The signal there is good.
Would you conclude that the signal is not merely good? / Would you conclude that the signal is not excellent?

C.1.2 HF16-Adapted items

Tunes for each item are shown in parentheses. Targets for the contrastive, non-contrastive, and unrelated conditions are shown in braces in that order.

1. {*painter, statue, register*}: Did the museum deliver any good news? They thrilled the sculptor (HLL).
Would you conclude that the museum did not thrill the painter?

2. {*doctor, clinic, plug*}: Did the murderer at the hospital strike again? He killed the nurse (LHSLL).
Would you conclude that the murderer did not kill the doctor?

3. {*dinosaurs, ice, corporate*}: Have the scientists dug anything up? They found fossilized mammoths (HLL).
Would you conclude that the scientists did not find fossilized dinosaurs?

4. {*building, river, interest*}: Did the engineer help with the city planning? He designed the bridge (LSHLH).
Would you conclude that the engineer did not design the building?

5. {*swan, nest, chain*}: Does Sammy like the animals at the park? He likes to feed the duck (HLH).
Would you conclude that Sammy does not like to feed the swan?

6. {*scarf, skinny, theory*}: Did the woman wear anything special to the party? She wore her favorite jeans (LHSLH).
Would you conclude that the woman did not want to wear her favorite scarf?

7. {*muffins, birthday, bracelet*}: Has the baker decided what dessert will be? He needs to make a cake (LSHLL).
Would you conclude that the baker did not need to make muffins?

8. {*jug, soup, lapel*}: Did the maid find the mouse? She found it in a can (LHSLL).
Would you conclude that the maid did not find the mouse in a jug?

9. {*tiara, posh, lecture*}: Did the model actually like any of the things she tried on? She adored the necklace (HLL).
   Would you conclude that the model did not adore the tiara?

10. {*puppy, furry, beach*}: Was it just me or did the kennel owner sound distracted on the phone? She was playing with a kitten (HLH).
    Would you conclude that the kennel owner was not playing with a puppy?

11. {*lawn, hoe, box*}: Did the farmer say if any of his crops were damaged in the storm? He checked on his garden (LHSLH).
    Would you conclude that the farmer did not check on his lawn?

12. {*houses, architects, weak*}: Did the tourists see anything interesting near the hotel? They saw historic buildings (LHSLL).
    Would you conclude that the tourists did not see historic houses?

13. {*train, baggage, splinter*}: Was the passenger able to get to his flight on time? He boarded the airplane (LSHLL).
    Would you conclude that the passenger did not board the train?

14. {*curving, line, gear*}: Did the manager have to take a detour to reach his appointment? The road he took was straight (HLH).
    Would you conclude that the road the manager took was not curving?

15. {*chair, dinner, pool*}: Is the craftsman selling anything at the flea market today? He built a table (LHSLL).
    Would you conclude that the craftsman did not build a chair?

16. {*cloudy, tropics, industry*}: Was Janet's wedding reception any fun? The day ended up rainy (LSHLL).
    Would you conclude that the day did not end up cloudy?

17. {*kind, slope, teal*}: Was this Emily's first time holding a newborn? She was very gentle (HLH).
    Would you conclude that the mother was not very kind?

18. {*economical, durable, organic*}: Did the shopper find any good sales on appliances at the store? She found them inexpensive (LHSLH).
    Would you conclude that the shopper did not find the prices economical?

19. {*emotional, mental, country*}: Was the victim okay when she heard the news? She became hysterical (HLH).
    Would you conclude that the victim did not became emotional?

20. {*physics, numbers, tooth*}: Do the students have anything due tomorrow? They have homework for math (LHSLL).
    Would you conclude that the students did not have homework for physics?

21. {*damp, rain, exile*}: Did you really run all the way to class without an umbrella? My blazer got all wet (HLH).
    Would you conclude that the blazer did not get damp?

22. {*sleet, frozen, blocks*}: Is the weather in Chicago bad in January? The city gets lots of snow (LHSLH).
    Would you conclude that the city does not get lots of sleet?

23. {*strange, even, potato*}: Did the fashion designer really make her own outfit for such a formal occasion? It was rather odd (LSHLL).
    Would you conclude that the outfit was not strange?

24. {*bathroom, stove, eagle*}: Did the girl ever find her missing shoe? It was in the kitchen (LHSLL).
    Would you conclude that the girl did not find her shoe in the bathroom?

25. {*pen, eraser, rake*}: Do the students know what to bring to the SAT? They were told to use a pencil (HLL).
    Would you conclude that the students were not told to write with a pen?

26. {*spotted, dull, fruit*}: Do the kittens look like their mother? All of them were striped (LSHLH).
    Would you conclude that the kittens were not brown?

27. {*toad, pond, jail*}: Did the toddler have fun in the backyard? He played with a turtle (LHSLH).
    Would you conclude that the toddler did not play with a toad?

28. {*oranges, sour, useful*}: Did the host make any cocktails for the party? He made one with lemons (LSHLL).
    Would you conclude that the host did not make a drink with oranges?

29. {*fish, slimy, kiss*}: Did the guest find anything hindden in the tank? She happened to notice the eel (LHSLL).
    Would you conclude that the guest did not happen to notice the fish?

30. {*church, priest, collie*}: Did the family check out any of the tour guide's recommendations? They visited the cathedral (LSHLH).
    Would you conclude that the family did not visit the church?

31. {*prosecutor, lawsuit, popcorn*}: Has your aunt gotten someone to handle the case yet? She hired a skilled attorney (HLL).
    Would you conclude that the aunt did not hire a skilled prosecutor?

32. {*bass, string, crutch*}: Is Jack going to join the youth orchestra this year? He wants to play the cello (LHSLL).
    Would you conclude that the boy did not want to play the bass?

33. {*purple, grass, tyranny*}: Did the junior class coordinate their outfits for the field trip? They all wore green (LSHLL).
Would you conclude that the junior class did not wear purple?

34. {*tulips, pink, tubas*}: Did Jane's fiancee get her anything nice for her birthday? He surprised her with roses (LHSLL).
Would you conclude that Jane's fiancee did not surprise her with tulips?

35. {*noodles, fried, patrol*}: Did the triplets ask for anything specific for dinner? They wanted to have rice (HLH).
Would you conclude that the triplets did not want to have noodles?

36. {*fork, lap, ears*}: Did the toddler have good manners at the dinner table? He asked for a napkin (LSHLH).
Would you conclude that the boy did not ask for a fork?

37. {*raincoat, raining, sword*}: Does the director have a jacket for the rain? He brought his umbrella (HLH).
Would you conclude that the director did not bring his raincoat?

38. {*pipe, lungs, cobra*}: Did the artist do anything after the dinner party? He enjoyed a cigarette (LSHLL).
Would you conclude that the artist did not enjoy a pipe?

39. {*circular, angles, experience*}: Have the homeowners settled on a layout for the baby's room? The newest room is square (HLH).
Would you conclude that the newest room was not circular?

40. {*leash, leather, planet*}: Did the family get their dog anything at the pet store? They got him a collar (LHSLL).
Would you conclude that they did not get their dog a leash?

41. {*cockroach, burrow, shingle*}: Has Michael had any issues with the new apartment? He realized it had termites (HLL).
Would you conclude that the apartment did not have cockroaches?

42. {*cabbage, salad, tournament*}: Did the fisherman get everything he needed from the market? He bought two pounds of lettuce (LSHLH).
Would you conclude that the fisherman did not buy cabbage?

43. {*sweet, chili, silly*}: Did the couple like the appetizer at the new restaurant? They thought it was too spicy (LHSLL).
Would you conclude that the couple did not think the appetizer was too sweet?

44. {*guilt, afraid, conceptual*}: Was the assistant prepared to talk with her boss? She was overcome with fear (HLL).
Would you conclude that the assistant was not overcome with guilt?

45. {*wide, length, cola*}: Did the couple like the driveway of the house they checked out? They thought it was too long (HLL).
Would you conclude that the couple did not think the driveway was too wide?

46. {*window, open, dip*}: Have the girls worked up the courage to sell cookies to the neighbors? They approached their door (LSHLL).
Would you conclude that the girls did not approach the window?

47. {*sunny, kite, cults*}: Was it peaceful living in the mountains for a year? It was always windy (HLL).
Would you conclude that the weather was not sunny?

48. {*ribbon, cotton, holy*}: Has the grandmother bought anything for her new project? She purchased some fabric (LSHLH).
Would you conclude that the grandmother did not purchase ribbon?

49. {*moon, bright, mop*}: Is it true the water was so still it was like a mirror? The lake reflected the sun (LHSLH).
Would you conclude that the lake did not reflect the moon?

50. {*mug, coffee, mole*}: Was the manager just doing dishes in the berakroom? He cleaned out his cup (LSHLL).
Would you conclude that the manager did not clean his mug?

51. {*brave, ballad, couch*}: Did the officer really save the day? His actions were heroic (LSHLH).
Would you conclude that the officer's actions were not brave?

52. {*toddler, embryo, wheel*}: Has the nanny made sure the kids are asleep? She checked on the baby (LHSLH).
Would you conclude that the nanny did not check on the toddler?

53. {*tent, woods, ash*}: Doesn't the family down the street rent out a property for the summer? They own a lake cabin (LSHLH).
Would you conclude that the family did not own a tent?

54. {*quake, clouds, zebra*}: The village is in bad shape, was it destroyed by some natural disaster? It was hit by a hurricane (HLL).
Would you conclude that the village was not hit by a quake?

55. {*snake, antler, truth*}: Did the farmer ever find any wild animals on his property? One time he found a deer (LSHLH).
Would you conclude that the farmer did not find a snake?

56. {*pepper, ocean, trees*}: Was the meat at the new steakhouse any good? It needed more salt (LSHLL).
Would you conclude that the meat did not need more pepper?

57. {*parasite, microscope, impulse*}: Did the doctors find out what caused the infection? My cut got some bacteria (LHSLH).
    Would you conclude that the infection was not caused by a parasite?

58. {*pears, ripe, polite*}: The workday is so long, do you bring any snacks? I always have apples (HLH).
    Would you conclude that he does not always have pears?

59. {*forest, water, cowboy*}: Does that off road path go anywhere interesting? It runs beside the swamp (LHSLH).
    Would you conclude that the path did not run beside the forest?

60. {*caviar, stream, roof*}: Did the weekend menu have anything special to offer? It featured salmon (LSHLH).
    Would you conclude that the menu did not feature caviar?

61. {*recorders, lenses, ice*}: Did the administration really let the press into the senator's funeral? They were told no cameras (LHSLH).
    Would you conclude that the press was not told no recorders?

C.1.3 Filler Items

Tunes for each item are shown in parentheses.

1. *woab*: Have you seen my keys anywhere? They're on the kitchen counter (LSHLH).
   Would you conclude that he has not seen the keys?

2. *strulk*: Did you do anything fun this weekend? I went hiking in the mountains (LHSLH).
   Would you conclude that he did not do anything fun?

3. *phirp*: Is there an electric car charging station around here? We have a gas station (HLH).
   Would you conclude that there is no charging station?

4. *sizz*: Did you do the extra readings for class? My printer was out of ink (LSHLH).
   Would you conclude that he did not do the readings?

5. *valc*: What's the weather supposed to be like today? It should be nice outside (HLH).
   Would you conclude that the weather will not be bad?

6. *hieb*: Have you done any traveling lately? Over break I went downtown (LHSLL).
   Would you conclude that he has not been travelling?

7. *rhymp*: Do you take cream with your coffee? I'll take almond milk (LHSLH).
   Would you conclude that he does not like black coffee?

8. *glive*: Can you play any instruments? I can play the guitar (HLH).
   Would you conclude that he can play other instruments?

9. *urnt*: Are you free this weekend? My parents are in town (HLH).
   Would you conclude that he is not free?

10. *sulte*: Did you watch the new avatar movie? I hated the first one (LSHLH).
    Would you conclude that he did not watch the movie?

11. *sloge*: Would you recommend the new italian restaurant? Their breadsticks are tasty (LSHLL).
    Would you conclude that he does not like the restaurant?

12. *nymb*: Could I borrow your calculus textbook? I left mine at home (HLH).
    Would you conclude that she can't borrow the textbook?

13. *lant*: Did anyone interesting present their research? I saw Jeremy (LSHLH).
    Would you conclude that jeremy did not present their research?

14. *porg*: Did you feed the animals? I fed some of them (LHSLH).
    Would you conclude that he did not feed all the animals?

15. *strope*: Have you taken organic chemistry yet? The first quarter is the worst (LSHLL).
    Would you conclude that he has not taken organic chemistry?

16. *shrusk*: Did you go to the beach for spring break? I don't like all the sand (HLH).
    Would you conclude that he did not go to the beach?

17. *gemf*: Did you eat breakfast this morning? I drank some coffee (LSHLH).
    Would you conclude that he did not eat breakfast?

18. *flin*: Do you want to get chinese for dinner? I have a paper due tonight (LHSLH).
    Would you conclude that he does not want chinese for dinner?

19. *demn*: Is it going to rain later? It's really cloudy outside (LSHLH).
    Would you conclude that it is not going to rain?

20. *fraik*: Do you watch reality TV shows? I like to watch the news (LHSLL).
    Would you conclude that he does not watch reality TV?

21. *mought*: Did you get tickets to the concert? I got into the queue (LHSLL).
    Would you conclude that he did not get the tickets?

22. *spote*: Did the boys come home late last night? They got back around two (LSHLL).
    Would you conclude that the boys did not come home late?

23. *zench*: Did you go to any of the recommended restaurants? I checked out a cafe (HLL).
    Would you conclude that he did not go to any of the recommended restaurants?

24. *mouge*: Do you have a phone I could use? The desk has a landline (LHSLH).
    Would you conclude that he does not have a phone she could use?

25. *snoog*: Were you in charge of planning the event? I catered the event (HLH).
    Would you conclude that he was not in charge of the event?

26. *smob*: Do you ever donate to charity? I give to st jude's hospital (LSHLH).
    Would you conclude that he does not give to charity?

27. *marce*: Did you remember to buy a helmet with your new bike? I already had one (HLL).
    Would you conclude that he did not buy a helmet?

28. *zautch*: Do you believe in magic? That stuff is for kids (LSHLL).
    Would you conclude that he does not believe in magic?

29. *jelch*: Do you go shopping after Thanksgiving dinner? My aunt brings me along (LSHLL).
    Would you conclude that he does not go shopping after Thanksgiving?

30. *rhalk*: Have you ever been in a play? My school had theatre class (LHSLH).
    Would you conclude that he has not been in a play?

31. *yurg*: Is the professor's class hard to get into? His class has twenty spots (HLL).
    Would you conclude that the professor's class is not hard to get into?

32. *barsh*: Do you have change for the bus? I use my ventra card (LSHLL).
    Would you conclude that he does not have change for the bus?

33. *skall*: Did you stay for the encore? They played my favorite song (LSHLL).
    Would you conclude that he did not stay for the encore?

34. *swoom*: Do you have a favorite snack at home? I like baking cookies (HLL).
    Would you conclude that he does not have a favorite snack?

35. *ruick*: Do you have a big family? I was one of seven kids (LHSLH).
    Would you conclude that he does not have a big family?

36. *promf*: Do you read a lot of books? I read a lot for my classes (LSHLH).
    Would you conclude that he does not read a lot of books?

37. *praite*: Do undergrads usually walk to campus? They don't get free transport (HLH).
    Would you conclude that the undergrads do not walk to campus?

38. *zote*: Are there any cheap restaurants around? There used to be a burger king (LHSLH).
    Would you conclude that there are no cheap restaurants around?

39. *shrint*: Are you a morning person? I have to have coffee (LHSLH).
    Would you conclude that he is not a morning person?

40. *pewve*: Do you have any tattoos? My mother would kill me (LHSLL).
    Would you conclude that he does not have any tattoos?

41. *terl*: Did you vote in the mayoral election? I'm not a resident (HLL).
    Would you conclude that he did not vote in the election?

42. *scoof*: Do you try to limit your screen time? All of my classes are online (LHSLL).
    Would you conclude that he does not try to limit screen time?

43. *palc*: Do you ever volunteer? Once a month I give blood (LSHLH).
    Would you conclude that he does not volunteer?

44. *fryck*: Have you ever ridden a motorcycle? They're too dangerous for me (HLH).
    Would you conclude that he has not ridden a motorcycle before?

45. *croush*: Do you speak more than one language? I took high school spanish (LSHLH).
    Would you conclude that he does not speak another language?

46. *knouth*: Do you invest any money? I have a roth IRA (HLL).
    Would you conclude that he does not invest any money?

47. *wrowse*: Do you like cilantro on your tacos? To me it tastes like soap (LHSLL).
    Would you conclude that he does not like cilantro?

48. *cerl*: Do you have any hobbies? I don't have the time (HLL).
    Would you conclude that he does not have any hobbies?

49. *snate*: Do you like watching marvel movies? There are too many (HLL).
    Would you conclude that he does not like marvel movies?

50. *zomp*: Are you going to the party this weekend? I'm bringing some chips (LHSLL).
    Would you conclude that he is not going to the party?

51. *cweese*: Do you want to go to the club tonight? I really don't like dancing (LHSLL).
    Would you conclude that he does not want to go to the club?

52. *fult*: Do you know a good math tutor? My friend is a math major (LSHLL).
    Would you conclude that he does not know a good math tutor?

53. *cwaim*: Does the office have a printer? You need a cable to connect (HLH).
    Would you conclude that the office does not have a printer?

54. *soatch*: Would you be able to store my bike while i'm away? I don't really have space (LSHLL).
    Would you conclude that he cannot store the bike?

55. *sybe*: Will we be able to see the stars at night? There's too much light pollution (LHSLH).
    Would you conclude that they cannot see the stars?

56. *brise*: Is there no entry fee for the event? They recommend donating (LSHLH).
    Would you conclude that there is not an entry fee?

57. *slown*: Do you have a tent for the camping trip? My dad has one at home (LSHLL).
    Would you conclude that he does not have a tent?

58. *krove*: Do you get charged for using an ATM? My bank reimburses me (LHSLL).
    Would you conclude that his bank does not charge ATM fees?

59. *screim*: Is the direct flight to New York expensive? It isn't worth the price (HLL).
    Would you conclude that the direct flight is not expensive?

60. *croog*: Would you want to go skydiving over break? I've always been scared of heights (HLL).
    Would you conclude that he does not want to go skydiving?

61. *plidd*: Did you hand out candy to the trick or treaters? My door bell rang constantly (LHSLL).
    Would you conclude that he did not hand out candy to trick or treaters?

C.2 Auditory materials details

Stimuli were recorded in a double-walled sound attenuating recording booth using a Shure SM27 microphone. After manual inspection of 4226 recordings, there were 3980 recordings of 130 sentences (65 critical items * 2 conditions) suitable for further analysis. These recordings were then force aligned at the phone level using the Montreal Forced Aligner (MFA, McAuliffe et al., 2017b) and loosely audited to correct for the syllable boundaries of the nuclear accented syllable. The recordings were additionally annotated with a modified ToBI annotation scheme, which primarily serves to annotate landmarks for the resynthesis targets. An example of one recording is shown in Figure C1 (p. 241).

C.2.1 Acoustic analyses of materials

This section presents descriptive analyses of the auditory materials. Of particular interest to this analysis is that there are a large number of recordings of the same six intonational tunes produced in 130 different utterances (number of unique words=116) with a variety of segmental contexts, stress patterns, and lengths. The analysis will start at the utterance level, then narrow in to the nuclear interval, then discuss the accentual peak within this interval. The raw utterances with superimposed averages are shown in Figure C2 (p. 242), where the nuclear interval is approximately halfway into

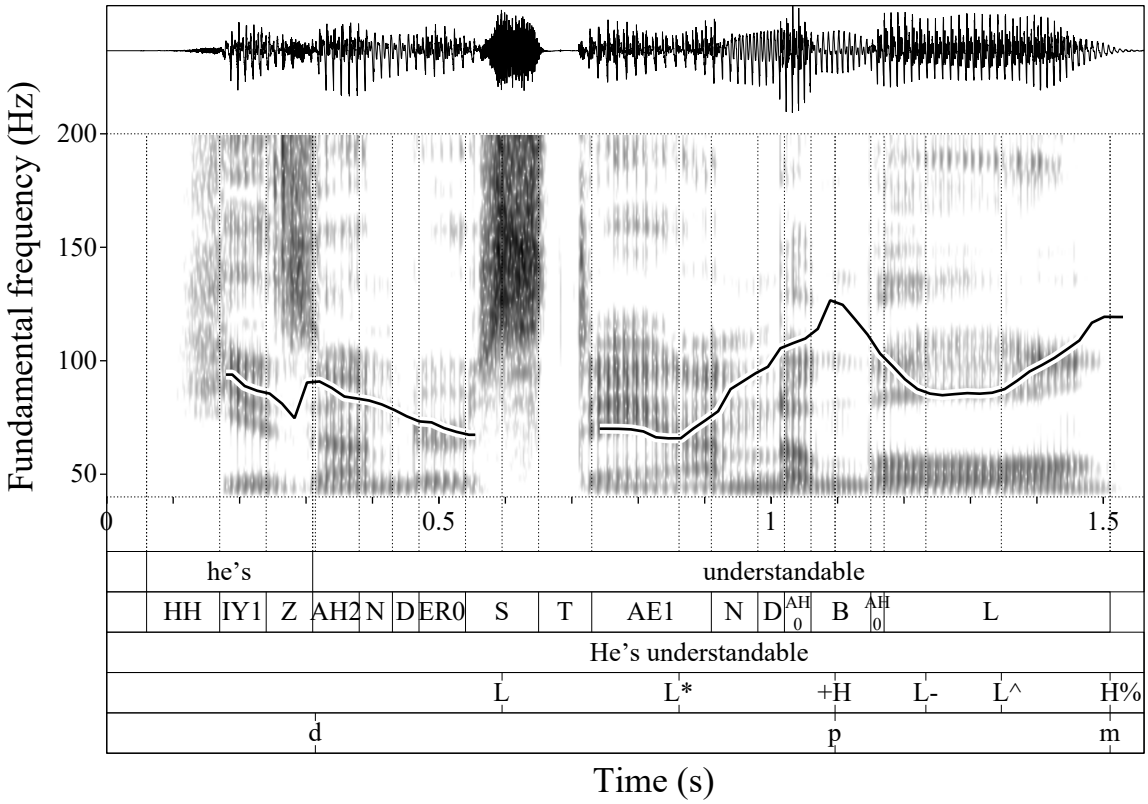


Figure C1: Annotation for a recording of *He's understandable* with L*+HL-H% intonation. The deviations from ToBI include `L` located roughly at the start of the low F0 interval for L*+H, `L*` located at the turning point for where the pitch accent onglide begins, `L-` marking the low target following the high accentual peak, and the use of `L^` marking where the final rise begins if there is a period of sustained low pitch (i.e., `L-` and `L^` together show primary and secondary association of the phrase accent). The bottom-most tier annotates where the prenuclear interval ends (d), the peak location (p), and the measured boundary tone target (m), which sometimes differed from the MFA utterance boundary.

the displayed contours.[1]

The nuclear intervals of the raw recordings are shown in Figure C3 (p. 243) both before and after landmark registration to the accentual peak. In other words, the contours are lined up based on the location of the peaks, ensuring that both the onglide to and offglide from the accentual peak can be modeled appropriately. In both cases, though, there is evidence of coarticulatory effects for L*+HL-H% where the L- target (the valley between the accentual peak and the boundary tone

[1]It should be noted that the averages presented in Figure C2 (p. 242) are not fully representative of the nuclear pitch contours because the peak locations are not aligned. Practically speaking, the lack of landmark registration has the results in underestimating the peak height (especially evident for L*+HL-L%) and flattening the peak (especially evident for H*L-L%)—see also discussion of alignment in Wang et al. (2016, p. 28). Regardless, from Figure C2 (p. 242), it can be seen that the H* is likely phonetically expressed with downstep given a prenuclear H*.

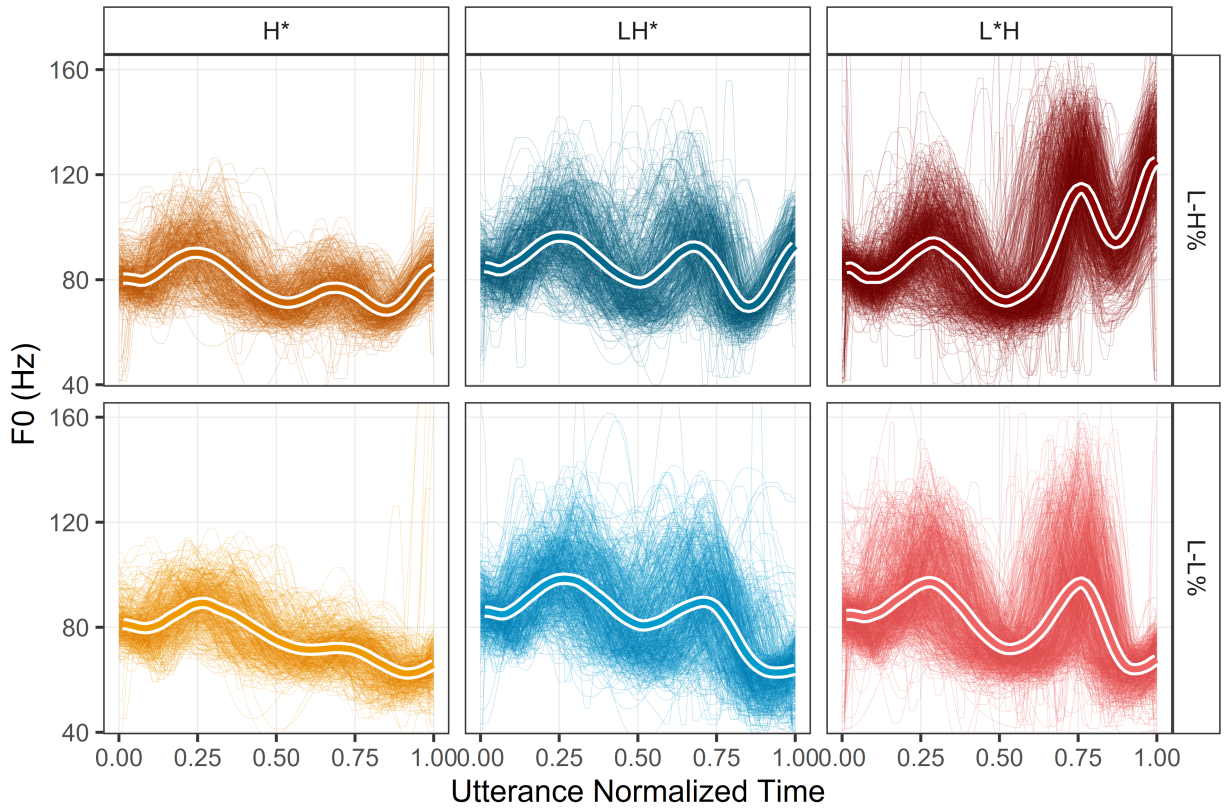


Figure C2: Raw F0 contours by tune with superimposed cross-sectional means. Colors are repeated in subsequent materials figures (e.g,. Figure C3 (p. 243)).

target) tends to be expressed higher than the speaker's floor (cf. the preceding L* target).[2]

Variation in the peak locations in the raw data is shown in Figure C4 (p. 244). Based on the figure, we can see that, generally, L*+H has higher and later peaks than both L+H* and H* and that L+H* generally has higher peak values than H*. This relationship is in line with previous descriptions of the three pitch accents (Iskarous et al., 2024).

[2]The coarticulatory effect for L*+HL-H% is a bit striking here but has not been (to my knowledge) formally described in prior work. However, a similar pattern can be seen in imitations of L*+HL-H% in Fig. 2 of Steffman et al. (2024, p. 10). Anecdotally, this coarticulatory effect is strongest when the nuclear pitch contour is used on a phrase final single-syllable word and weaker when there are additional syllables between the nuclear accented syllable and the intonational phrase boundary. Additionally, attempts to suppress this behavior in production by forcing F0 to return to the floor yields particularly unnatural productions of single-syllable words.

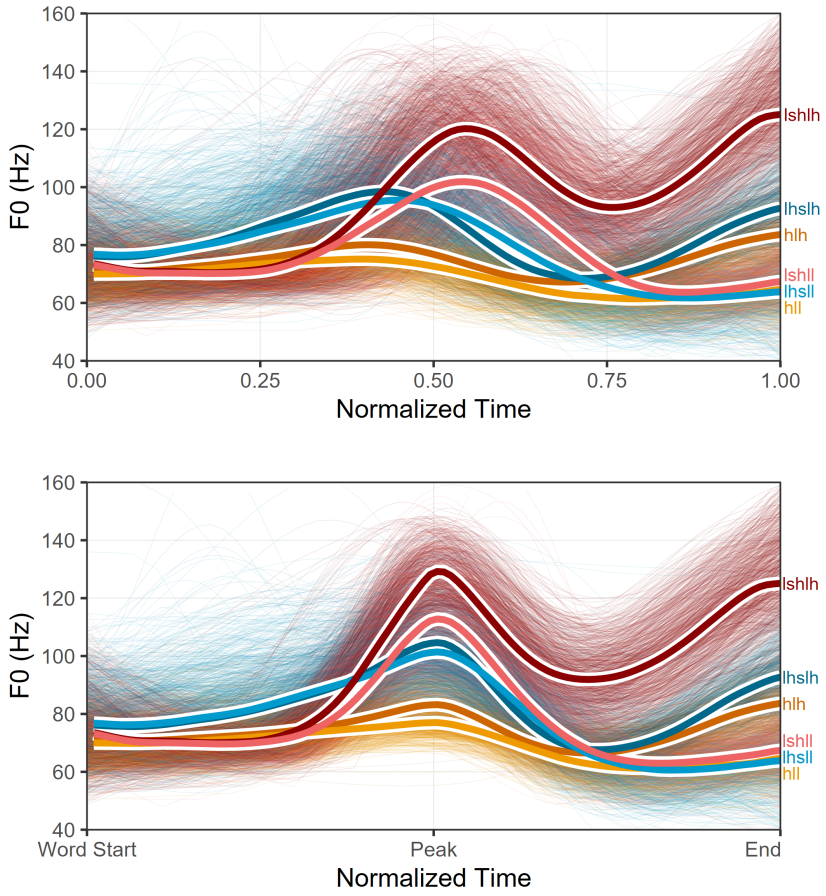


Figure C3: Raw pitch contours time-normalized in one of two ways. The left panel shows time normalization from the start to the end of the nuclear interval (=a single word). The right panel shows time normalization from the start of the nuclear accented word (left), up to the annotated accentual peak (middle), then up to the final F0 measurement at the end of the word (right). Raw contours are smoothed using splines for the sake of visualization only, but such smoothing is not part of the analysis or modeling of the contours.

C.2.1.1 Analysis of Peak Alignment and Height

Recent work has found robust evidence that the rising pitch accents for MAE follow a cline of both alignment and peak height such that pitch accents aligned later also rise to higher peak values (i.e., H* < L+H* < L*+H, Iskarous et al., 2024; Steffman et al., 2024). Although the landmark-registered contours are modeled using GAMMs (which will account for variation in peak height and onglide shape), the landmark registration, by necessity, removes the variation in peak alignment. Accordingly, a separate analysis of the peak alignment is provided in this section.

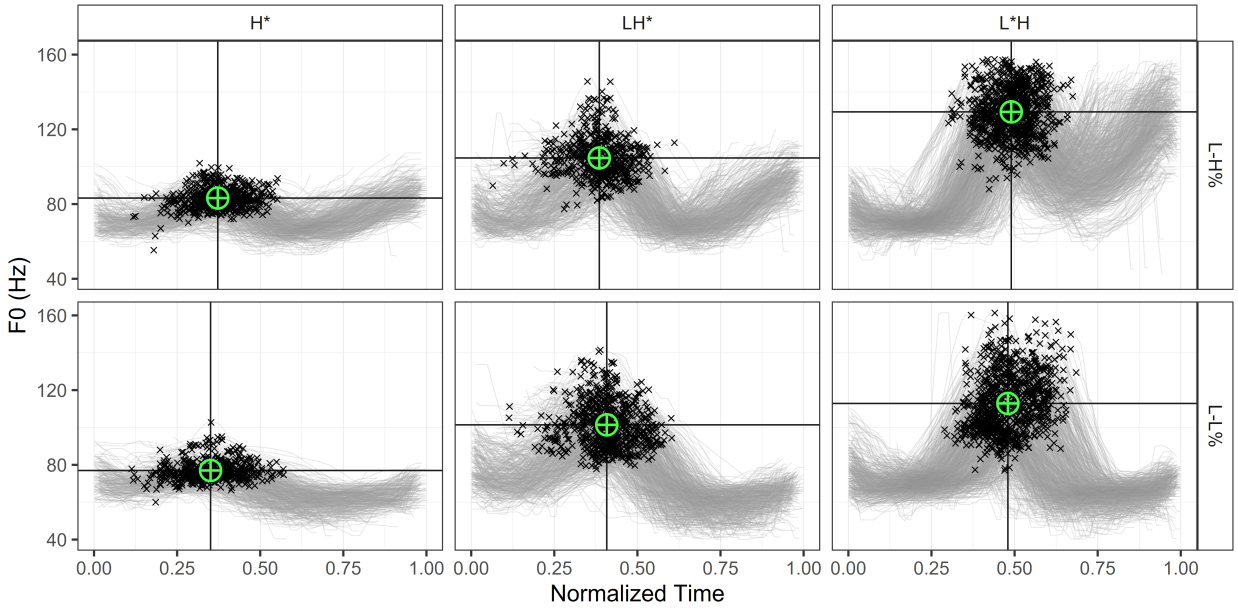


Figure C4: Raw pitch contours broken up by tune. Peak locations are annotated with an $\times$, while the crosshairs show the average location and height of the peak.

Peak location is measured in relative terms as the proportion of the stressed syllable duration. For example, an accentual peak occurring at the end of the stressed syllable would have a value of 100%; earlier alignment would fall between 0% and 100%; and later alignment would be beyond 100%. Because the nuclear accented words in these sentences have varying lengths and metrical patterns, the items are grouped into three stress groups: words with final stress (e.g., *cóld*, *obése*), words with penultimate stress (e.g., *tíny*, *mediócre*), and words with antepenult stress or earlier (e.g., *pálatable*, *understándable*).

The model includes predictors of pitch accent (H*, L+H*, L*+H), edge tone (L-L%, L-H%), and metrical group (Final, Penult, Antepenult) as well as the 2- and 3-way interactions between the three factors. The random effects structure additionally includes random intercepts by word and random slopes of pitch accent by word.[3] The model predictions are shown in Figure C5 (p. 245).

The full model results are reported in Table C1 (p. 248). Coefficients are interpretable on the percentage point scale (e.g., a difference of 0.15 suggests a 15 percentage point increase in align-

[3]Edge tone is scaled sum coded ($\pm.5$) with L-H% as the reference level. Both pitch accent and metrical group are backwards-difference coded, encoding two pairwise comparisons between adjacent levels.

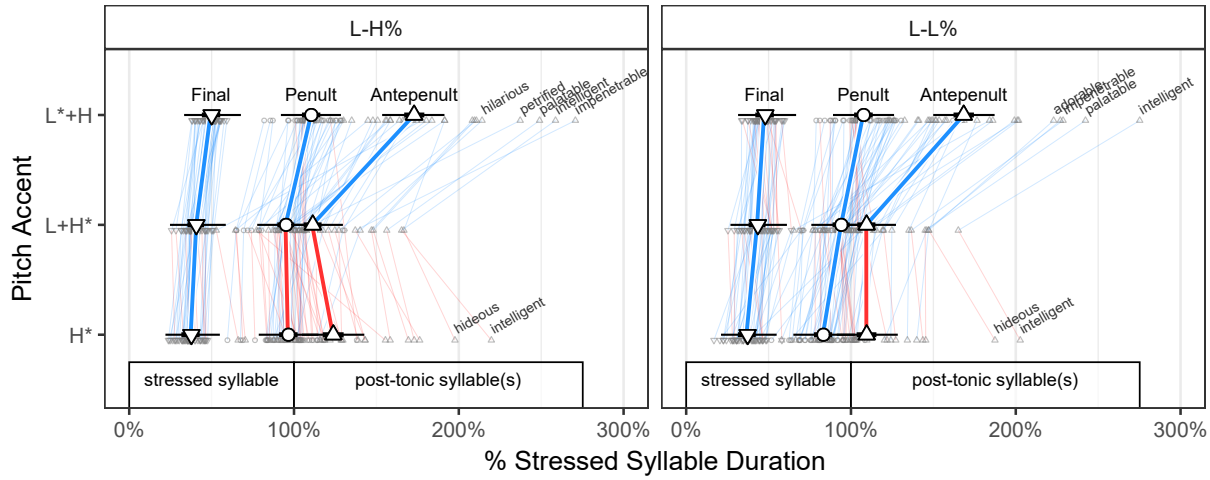


Figure C5: Posterior mean predictions for the peak alignment for each tune, broken up by edge tone, pitch accent, and metrical grouping. Comparing peak alignment of H* with L+H*, and L+h* with L*+H, positive alignment differences (i.e., later alignment) are shown in blue, while negative differences (i.e., earlier alignment) are shown in red. Lines show 95% credible intervals about the mean.

ment). The model shows that, overall, there is no credible difference (i.e., there is overlap) in the distribution of peak alignment for H* and L+H* ($\hat{\beta} = 0.01, CrI = [-0.01, 0.03]$) but that L*+H is credibly aligned later than L+H* ($\hat{\beta} = 0.27, CrI = [0.24, 0.3]$). The group with final stress is aligned earlier than the group with penultimate stress ($\hat{\beta} = -0.55, CrI = [-0.63, -0.48]$) and the penultimate group is aligned earlier than the antepenult group ($\hat{\beta} = -0.35, CrI = [-0.43, -0.27]$); these results reflect tonal compression effects when there are fewer syllables available. Notably, there appear to be some outliers in Figure C5 (p. 245) with alignment values that are very late. However, these words tend to have lax stressed vowels (*intelligent, impenetrable, petrified*), which are known to be shorter than tense vowels (i.a. Hillenbrand et al., 1995; Leung et al., 2016), which may have inflated the alignment of these targets.

C.2.1.2 GAMM Modeling

Variation in the time-normalized contours in the bottom panel of Figure C3 (p. 243) are modeled using a GAMM (Wood, 2017). The model predicts F0 (in Hz, as there is only one speaker) using fixed predictors of tune (a six-level factor) and metrical structure (an 11-level factor), along with

smooth terms by tune and metrical structure using basis splines with 9 knots and random smooths by word. The goal with this model is not statistical comparison between the two tunes, but rather to create targets to use for resynthesis. The GAMM-predicted contours for each tune are shown in Figure C6 (p. 246), where the expected variation in peak height across the three pitch accents is present. The L+H* trajectories (in blue) have a more domed onglide while the L*+H contours (red, pink) have a more scooped onglide. Recall that the H* contours (yellow, orange) are downstepped relative to their prenuclear region, so while the onglide appears to start from the same low point as the L+H* contours, this is likely more indicative of a shallow rising pitch excursion with a sagging transition (Ladd & Schepman, 2003).

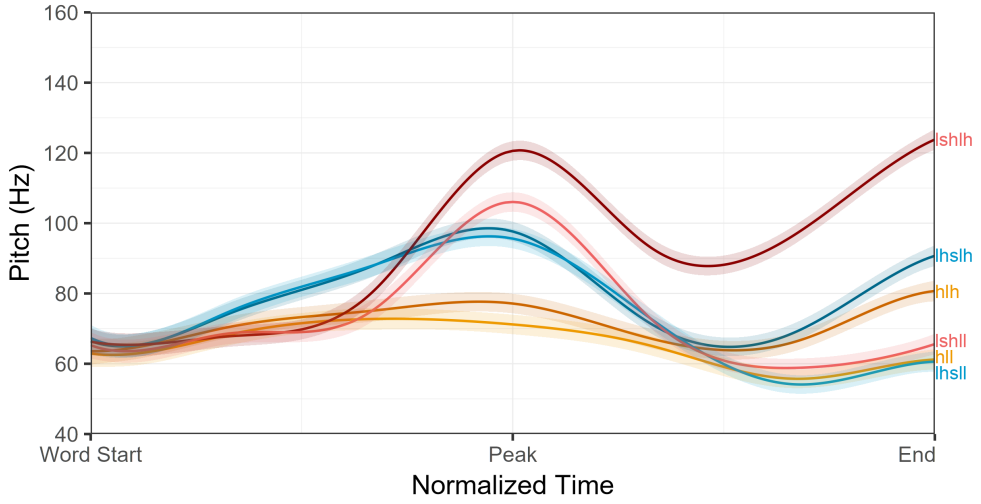


Figure C6: GAMM predictions for each of the six tunes, collapsing across metrical structure and word.

C.2.2 Resynthesis of Materials

The resynthesis targets are based on the GAMM results reported above, but slightly modified to make the L+H* onglide more distinct and to ensure that tunes ending in L-L% all had final falling trajectories (rather than a very slight rise, which was an artifact of the recording procedure). The onglide for the H* trajectory is intentionally not specified so that the onglide to the accentual peak is a rather gradual linear interpolation from the end of the prenuclear region. The onglide start, accentual peak, L- locations, and boundary tone locations are aligned in a piecewise fashion to the

ToBI landmark annotations described in Figure C1 (p. 241). Of the 3980 files, 3520 were deemed as potential candidates for resynthesis. The resynthesis quality of all files was manually checked to select the 780 (65 items$\times$2 conditions$\times$6 tunes) final recordings. The final resynthesized materials are shown in the main text in Figure 4.5 (p. 141). Compared to the original raw recordings in Figure C3 (p. 243), we can see that the amount of variation is greatly reduced.

The prenuclear region is scaled to a max peak height (wherever it occurs) to 90Hz for all tunes so as to minimize potential covarying cues that may cue the intended tune earlier in the utterance; Figure C7 (p. 248) in Appendix C.2 (p. 240) shows the final resynthesized utterances including the prenuclear region. The systematicity of anticipatory cues to nuclear tunes in the prenuclear region is beyond the scope of this work, but may well be a potential avenue for future work (see also Petrone & Niebuhr, 2014).

Recall that the intended materials are question-answer pairs, where the answers are the resynthesized utterances described in this section. The final resynthesized answers are then trimmed to avoid extraneous silence and are concatenated to the recorded questions (which are not manipulated via resynthesis) with an inter-speaker gap of 338ms.[4]

To summarize the acoustic materials, a large corpus of 3980 recordings of 130 utterances recorded with six intonational tunes was recorded. These recordings were analyzed to verify that they were representative of prior descriptions of the rising accents. The modeling results are then used to minimize the phonetic variation via resynthesis. The final set of resynthesized utterances was manually checked to select the most successful 780 resynthesized recordings. The filler items were similarly resynthesized based on the critical item trajectories.

Figure C7 (p. 248) shows the final resynthesized recordings including the prenuclear region. Note that these contours are time normalized using the entire utterance duration.

Table C1 (p. 248) shows the full modeling results.

[4]This gap length was based on taking the average of the results from (Levinson & Torreira, 2015, p. 9), who report 100ms turn latencies with no inbreaths and upwards of 576ms when inbreaths are present. With these materials, the former was too short while the latter was too long.

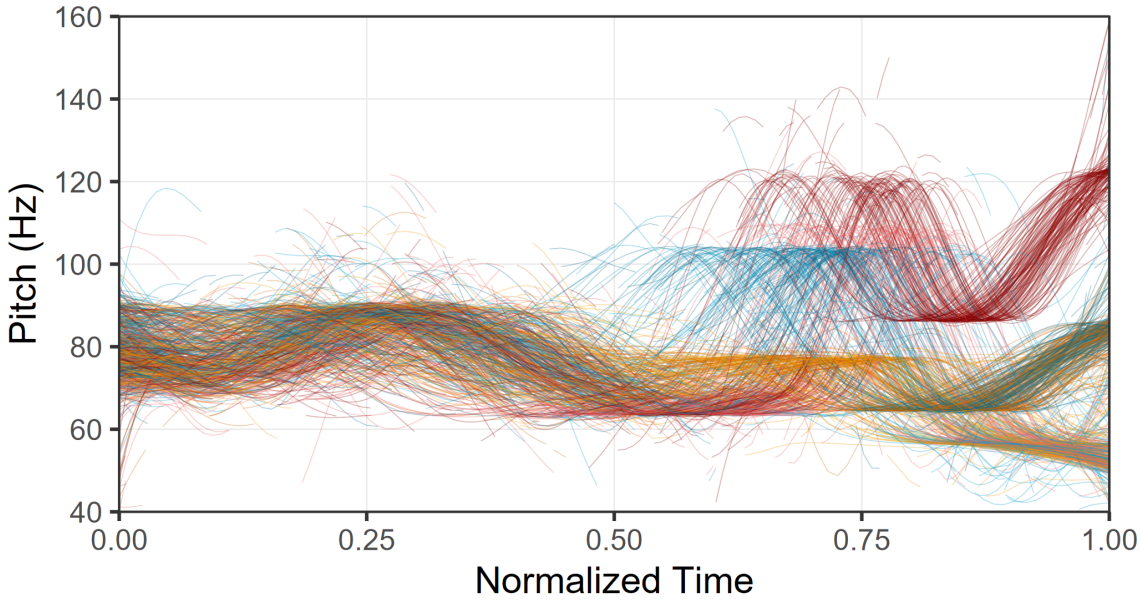


Figure C7: Final 780 resynthesized recordings, time normalized to the utterance duration. Note that these contours are not aligned at the accentual peak.

Term	Estimate	Std.Error	95% CrI	PD
Intercept	0.91	0.02	[0.88, 0.94]	100.00
LH*−H*	0.01	0.01	[−0.01, 0.03]	86.17
L*H−LH*	0.27	0.02	[0.24, 0.30]	100.00
Penult−Antepenult	−0.35	0.04	[−0.43, −0.27]	100.00
Final−Penult	−0.55	0.04	[−0.63, −0.48]	100.00
LL−LH	−0.04	0.01	[−0.05, −0.03]	100.00
LH*:Penult	0.11	0.02	[0.06, 0.16]	100.00
L*H:Penult	−0.46	0.04	[−0.54, −0.38]	100.00
LH*:Final	0.00	0.02	[−0.04, 0.04]	50.71
L*H:Final	−0.08	0.04	[−0.15, −0.01]	98.28
LH*−H*:LL	0.09	0.01	[0.07, 0.12]	100.00
L*H−LH*:LL	−0.03	0.01	[−0.05, −0.01]	99.74
Penult:LL	0.01	0.02	[−0.02, 0.04]	77.65
Final:LL	0.06	0.01	[0.03, 0.09]	100.00
LH*:Penult:LL	0.00	0.04	[−0.07, 0.07]	51.52
L*H:Penult:LL	0.01	0.03	[−0.05, 0.06]	58.90
LH*:Final:LL	−0.10	0.03	[−0.15, −0.04]	99.93
L*H:Final:LL	−0.03	0.02	[−0.07, 0.02]	90.40

Table C1: Statistical model summary for the peak-alignment model of the Chapter 4 materials.

C.3 Norming Task Item Breakdown

Figure C8 (p. 250) and Figure C9 (p. 251) show the by-item SI and MI rates, respectively, for the norming task. Figure C10 (p. 252) shows the by-item acceptability ratings for the norming task.

C.4 Exp. 1 (Auditory SI Task) Details

The contrast matrix for the logistic regression model does not correspond to typical contrast schemes (see documentation in Sostarics, 2024 for a review). This matrix is shown in Table C2 (p. 249).

	lhsll-hll	lshll-hll	RFR-Fall	lhslh-hlh	lshlh-hlh
hll	-1/3	-1/3	-1/2	0	0
lhsll	2/3	-1/3	-1/2	0	0
lshll	-1/3	2/3	-1/2	0	0
hlh	0	0	1/2	-1/3	-1/3
lhslh	0	0	1/2	2/3	-1/3
lshlh	0	0	1/2	-1/3	2/3

Table C2: Custom contrast matrix encoding pairwise comparisons from H* within each broad tune class, with an overall comparison across RFR and Fall tune classes. Column names indicate comparisons while row names indicate the tune levels in a shortened notation (e.g., lhsll = L+H*L-L%).

The by-item SI rates are broken up into two figures for falls (Figure C11 (p. 253)) and RFR-shaped tunes (Figure C12 (p. 254). Note that the ordering of the Y-axis matches the order shown for the norming task results in Figure C8 (p. 250).

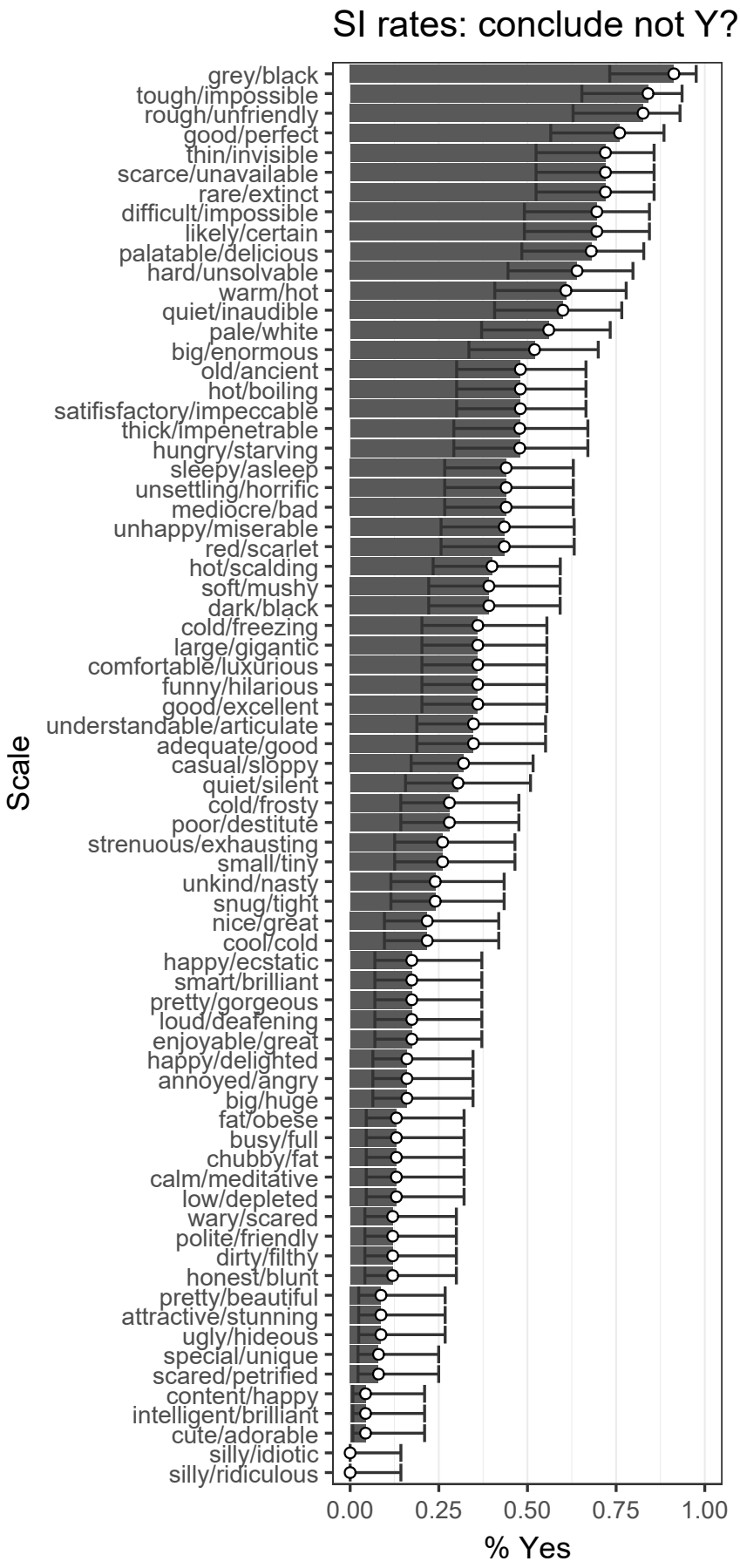


Figure C8: By-item response rates (% Yes) to the SI-probing questions in the norming task, e.g., *Would you conclude that the office is not cold?*. Error bars reflect 95% Wilson score intervals.

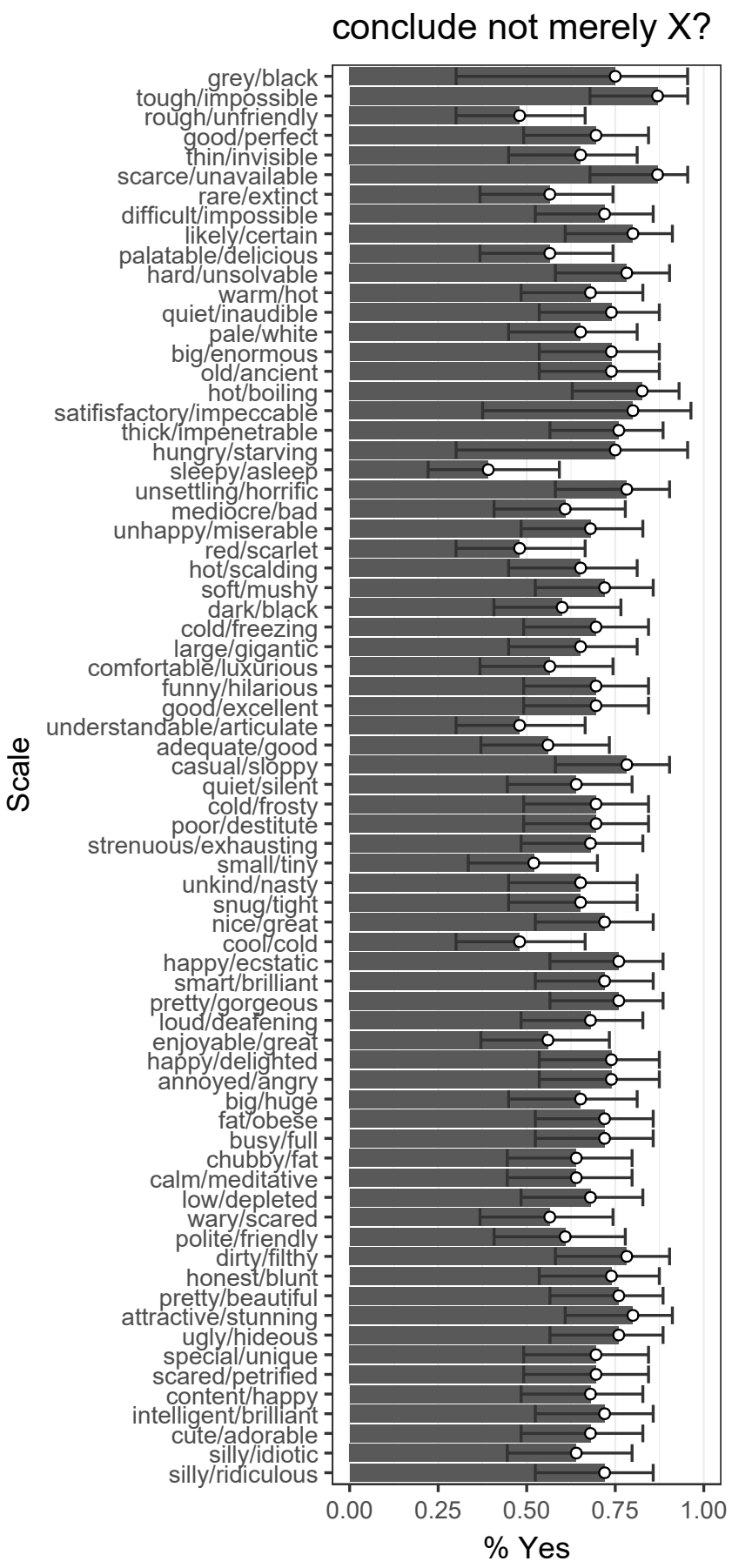


Figure C9: By-item response rates (% Yes) to the MI-probing questions in the norming task, e.g., *Would you conclude that the office is not merely cool?*. Scales are ordered as in Figure C8 (p. 250). Error bars reflect 95% Wilson score intervals.

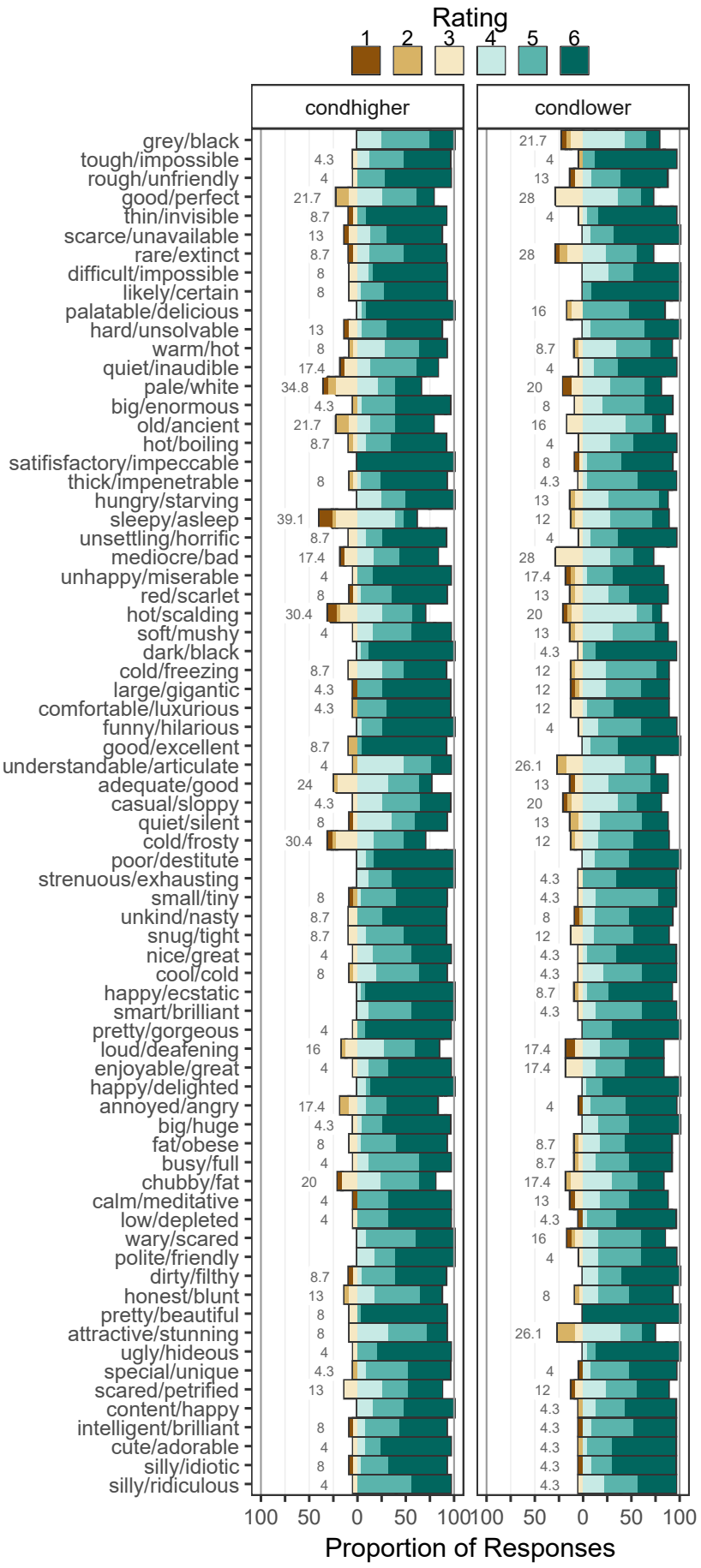


Figure C10: Empirical acceptability ratings when the higher scalemate is used (left) and when the lower scalemate is used (right), shown as the proportion of ratings for each item. The bars are centered on the midpoint of the scale, with lower ratings shown to the left in yellow and higher ratings shown to the right in green. Numbers on the left show the total proportion of low (rating=1-3) ratings.

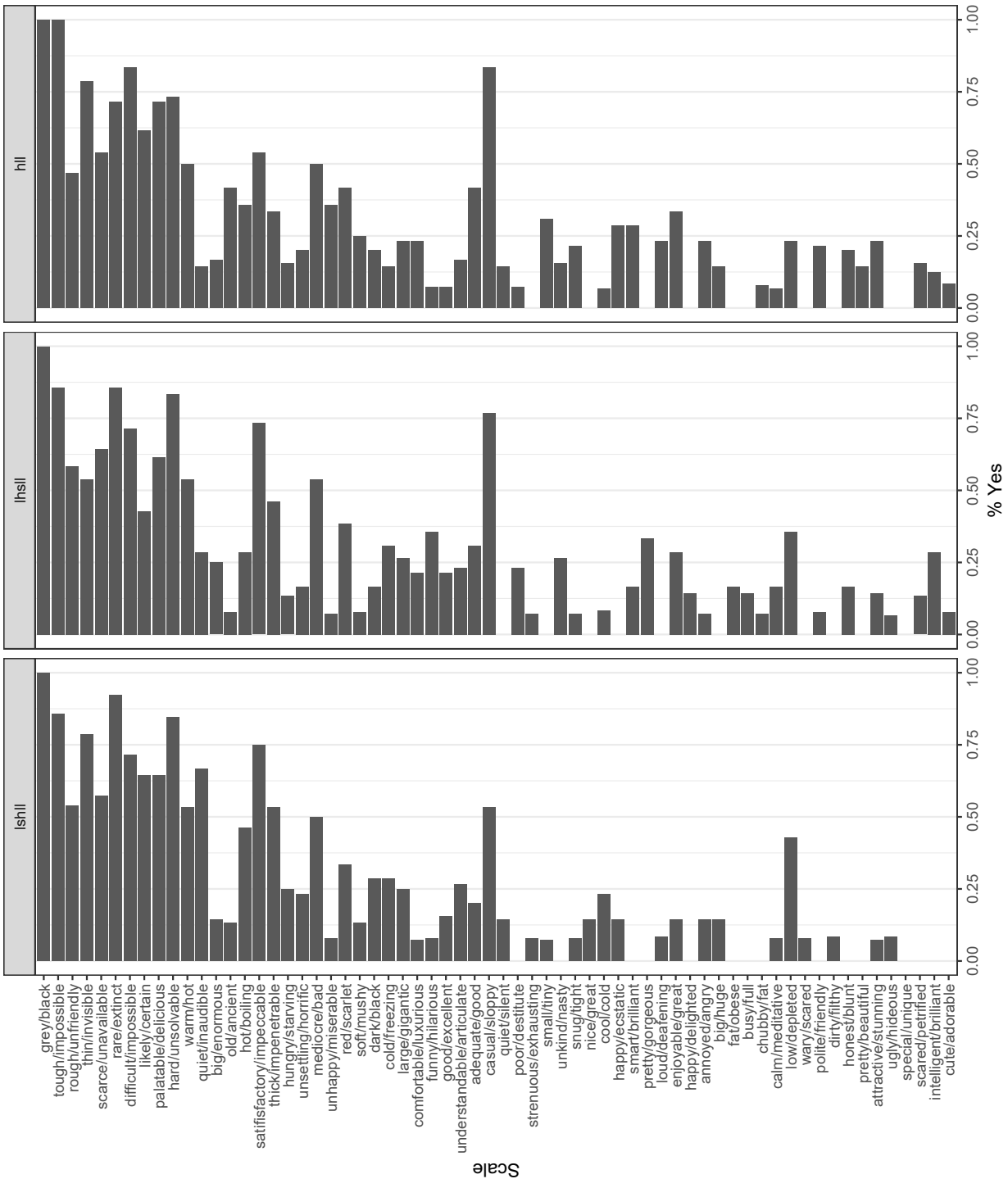


Figure C11: By-item SI rates for Falling tunes in Exp. 1 (web-based auditory inference task).

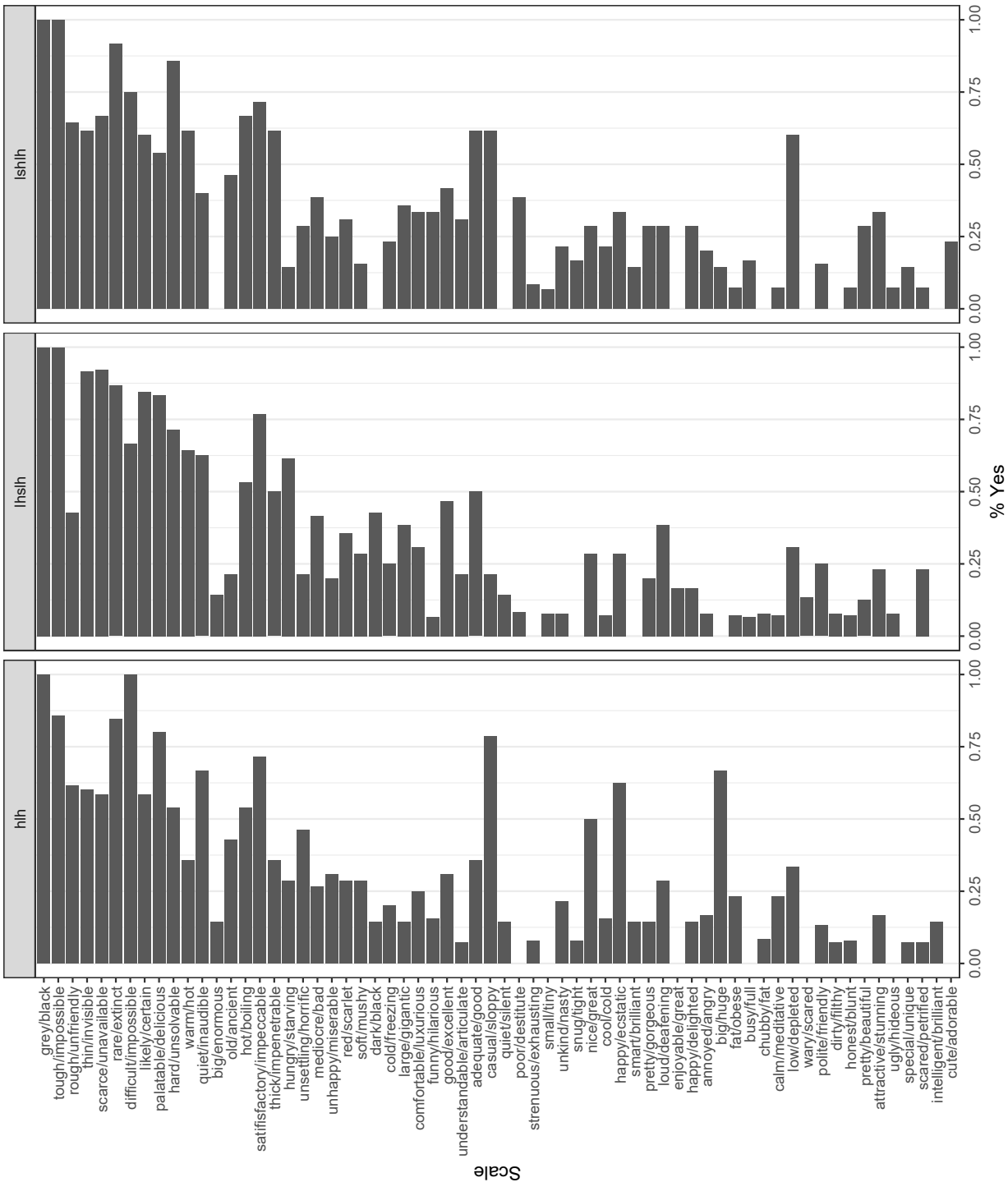


Figure C12: By-item SI rates for RFR-shaped tunes in Exp. 1 (web-based auditory inference task).

C.5 Exp. 2 (In-Person 750ms SOA Lexical Decision)

Regarding the model structure in Experiment 2, there are a few unorthodox design decisions which make answering the main questions straightforward but are nonetheless worth noting. First, the five-level Condition predictor is Helmert coded but sets the intercept at the mean of the HIGHERTARGET condition, rather than the typical grand mean; the contrast matrix for this is shown in Table C3 (p. 255). When combined with the sum-coded Tune predictor, the deviations encoded by each Tune comparison are no longer deviations from the grand mean (as is typical for sum coding) but rather encode deviations from the HIGHERTARGET mean. Going further, this combination of contrasts yields interaction terms whose interpretation is the difference between the logRT in the lower condition minus the RT in the higher condition.[5]

	Higher −Lower	Contrastive −Scalar	NonContrastive −Contrastive	Unrelated −Related
LowerTarget	-1	0	0	0
HigherTarget	0	0	0	0
Contrastive	-1/2	1	0	0
NonContrastive	-1/2	1/3	1	0
Unrelated	-1/2	1/3	1/4	1

Table C3: Contrast matrix for the Condition predictor. Row names indicate condition levels while column names indicate individual comparisons. The intercept is set at the HigherTarget condition.

The other thing worth noting is that the analysis presented here differs slightly from that presented in Sostarics et al. (2025). There, because the HF16-adapted items were presented with only one intonational tune, the comparisons between conditions and the comparisons between tunes within the critical conditions were split into two models. The model here combines all the data into one model, which helps to incorporate additional information about the participants' distributions of RT values to real-word targets and also allows for the effects of word length, frequency, and block to be more representative of a larger set of words. Ultimately though, the differences between the models do not change the pattern of results. The full model results are reported in

[5]To give an example in terms of percent change, if the $\%\Delta$ for H*L-H% is -1.96% for the HigherTarget condition and +0.45% in the LowerTarget condition, then the interaction term reflects $-1.96 - 0.45 = -2.41$.

Table C4 (p. 257).

C.5.1 Exp. 2b (Web-based Lexical Decision) Experiment Details

This section provides further details about the lexical decision task administered on Prolific. The distribution of reaction times (RT) and response durations (RD) are shown in Figure C13 (p. 260). RTs slower than 6 seconds or beyond 3.5 times the interquartile range above the participant's median log RT are excluded, resulting in a loss of 0.801% of the data. Note that RT measurements are not excluded on the basis of the RD values for the RT analyses presented here, but the pattern of results is not affected in either case.

The keen-eyed reader may notice what appears to be horizontal bands in the web-based experiment data shown in Figure C13 (p. 260), suggesting that there is periodicity present in the RD measurements. The eagle-eyed reader may notice a similar pattern vertically, suggesting a similar periodicity in the RT measurements as well. These observations were confirmed via a cepstral analysis, shown in Figures C14 (p. 261) and C15 (p. 262), which showed that both RDs and RTs in the web-based experiment show a 60Hz periodicity (i.e., higher density at values $\approx$16.67 milliseconds apart). This pattern is more robust for the RD measurements (Cepstral Peak Prominence (CPP)=0.38 at 16.74ms) than the RT measurements (CPP=0.19 at 16.92ms). The in-person data (which was better controlled and used better hardware) does not show evidence of such patterns (Figure C15 (p. 262)). To the best of my knowledge, this is not an issue with PsychoJS specifically but likely an idiosyncratic web browser issue involving event synchronization to the monitor refresh rate that is beyond the control of the experiment implementation. The takeaway here is that the web-based experiment RT measurements are contaminated, which substantially complicates identifying subtle effects of intonation.

The posterior predicted RTs for the individual target conditions are shown in Figure C16 (p. 263). The distribution of posterior predicted percent change values are shown in Figure C17 (p. 263). The model results are shown in Table C5 (p. 259). The main results of interest are that the effect of semantic relatedness ($\hat{\beta} = 0.053, CrI = [0.026, 0.080], PD = 99.98$)

Term	Estimate	%Δ	SE	95% CrI		PD
HigherTarget Mean	6.226		0.019	[6.189,	6.262]	100.00
WordLogFrequency	−0.031	−3.01	0.003	[−0.036,	−0.025]	100.00
WordLength	0.007	0.75	0.002	[0.003,	0.012]	99.87
Block	−0.011	−1.07	0.001	[−0.013,	−0.008]	100.00
H*LH	−0.020	−1.95	0.008	[−0.035,	−0.004]	99.30
LH*LH	0.005	0.46	0.008	[−0.011,	0.021]	71.56
LH*LL	−0.010	−1.02	0.008	[−0.026,	0.006]	89.83
L*HLH	0.019	1.94	0.008	[0.003,	0.035]	99.14
L*HLL	0.006	0.62	0.008	[−0.010,	0.022]	77.49
Higher−Lower	−0.005	−0.50	0.012	[−0.028,	0.018]	66.69
NonScalar−Scalar	0.007	0.67	0.012	[−0.017,	0.030]	71.35
Ncont−Contrastive	0.043	4.42	0.014	[0.016,	0.070]	99.85
Unrelated−Related	0.084	8.74	0.013	[0.059,	0.108]	100.00
Higher−Lower						
H*LH	−0.022	−2.19	0.011	[−0.044,	0.000]	97.47
LH*LH	0.004	0.37	0.011	[−0.018,	0.026]	62.60
LH*LL	−0.014	−1.39	0.011	[−0.036,	0.009]	89.28
L*HLH	0.019	1.90	0.011	[−0.003,	0.041]	95.42
L*HLL	0.010	0.99	0.011	[−0.012,	0.032]	80.63
NonScalar−Scalar						
H*LH	−0.006	−0.62	0.017	[−0.040,	0.028]	64.15
LH*LH	0.021	2.15	0.018	[−0.014,	0.057]	88.12
LH*LL	−0.009	−0.88	0.017	[−0.043,	0.025]	69.41
L*HLH	−0.003	−0.29	0.018	[−0.037,	0.032]	57.04
L*HLL	−0.008	−0.79	0.018	[−0.044,	0.029]	67.12
Ncont−Contrast						
H*LH	0.011	1.12	0.022	[−0.032,	0.053]	69.98
LH*LH	−0.002	−0.15	0.022	[−0.044,	0.041]	52.51
LH*LL	0.009	0.95	0.019	[−0.029,	0.047]	69.05
L*HLH	0.002	0.23	0.021	[−0.039,	0.044]	54.35
L*HLL	−0.012	−1.17	0.020	[−0.051,	0.027]	72.14
Unrelated−Related						
H*LH	0.006	0.57	0.019	[−0.032,	0.042]	62.57
LH*LH	−0.019	−1.89	0.018	[−0.054,	0.018]	85.49
LH*LL	−0.005	−0.47	0.018	[−0.041,	0.032]	60.35
L*HLH	−0.011	−1.14	0.019	[−0.049,	0.027]	73.40
L*HLL	0.019	1.94	0.019	[−0.019,	0.056]	84.62

Table C4: Exp. 2 (long SOA) full model summary.

and the effect of contrastive associates versus non-contrastive associates ($\hat{\beta} = 0.025, CrI = [0.000, 0.050], PD = 97.63$) were replicated. However, further distinctions within the contrastive associates are not credibly different.

Term	Estimate	%Δ	SE	95% CrI	PD
HigherTarget Mean	6.565		0.035	[6.496, 6.634]	100.00
WordLogFrequency	−0.029	−2.89	0.003	[−0.034, −0.024]	100.00
WordLength	0.005	0.48	0.002	[0.000, 0.009]	98.28
Block	−0.012	−1.17	0.001	[−0.015, −0.009]	100.00
H*LH	0.005	0.51	0.009	[−0.012, 0.023]	71.13
LH*LH	0.006	0.63	0.009	[−0.012, 0.025]	74.96
LH*LL	−0.011	−1.08	0.009	[−0.029, 0.007]	88.83
L*HLH	−0.005	−0.51	0.009	[−0.023, 0.013]	71.34
L*HLL	−0.011	−1.06	0.009	[−0.028, 0.007]	88.04
Higher−Lower	0.000	0.00	0.011	[−0.021, 0.022]	50.27
NonScalar−Scalar	0.004	0.38	0.011	[−0.018, 0.026]	63.44
Ncont−Contrastive	0.025	2.51	0.013	[0.000, 0.050]	97.63
Unrelated−Related	0.053	5.46	0.014	[0.026, 0.080]	99.98
Higher−Lower					
H*LH	−0.001	−0.10	0.013	[−0.026, 0.024]	53.31
LH*LH	0.001	0.09	0.013	[−0.024, 0.025]	52.76
LH*LL	−0.015	−1.48	0.012	[−0.039, 0.009]	88.60
L*HLH	−0.013	−1.27	0.013	[−0.037, 0.013]	84.05
L*HLL	0.000	−0.04	0.013	[−0.025, 0.024]	51.07
NonScalar−Scalar					
H*LH	−0.004	−0.39	0.019	[−0.042, 0.035]	57.97
LH*LH	0.019	1.87	0.018	[−0.018, 0.053]	84.27
LH*LL	−0.006	−0.61	0.017	[−0.040, 0.027]	63.82
L*HLH	−0.003	−0.25	0.018	[−0.038, 0.033]	55.30
L*HLL	−0.004	−0.42	0.018	[−0.040, 0.031]	59.38
Ncont−Contrast					
H*LH	0.014	1.46	0.020	[−0.025, 0.054]	76.02
LH*LH	−0.008	−0.78	0.019	[−0.046, 0.030]	66.12
LH*LL	0.016	1.59	0.019	[−0.022, 0.053]	79.87
L*HLH	−0.003	−0.33	0.020	[−0.043, 0.036]	56.69
L*HLL	−0.018	−1.82	0.018	[−0.054, 0.019]	84.17
Unrelated−Related					
H*LH	0.002	0.20	0.020	[−0.038, 0.042]	54.07
LH*LH	−0.024	−2.33	0.021	[−0.063, 0.017]	87.21
LH*LL	−0.008	−0.81	0.021	[−0.048, 0.033]	65.83
L*HLH	0.007	0.67	0.021	[−0.033, 0.047]	63.18
L*HLL	0.001	0.13	0.021	[−0.041, 0.043]	52.95

Table C5: Exp. 2b (web-based lexical decision) full model summary.

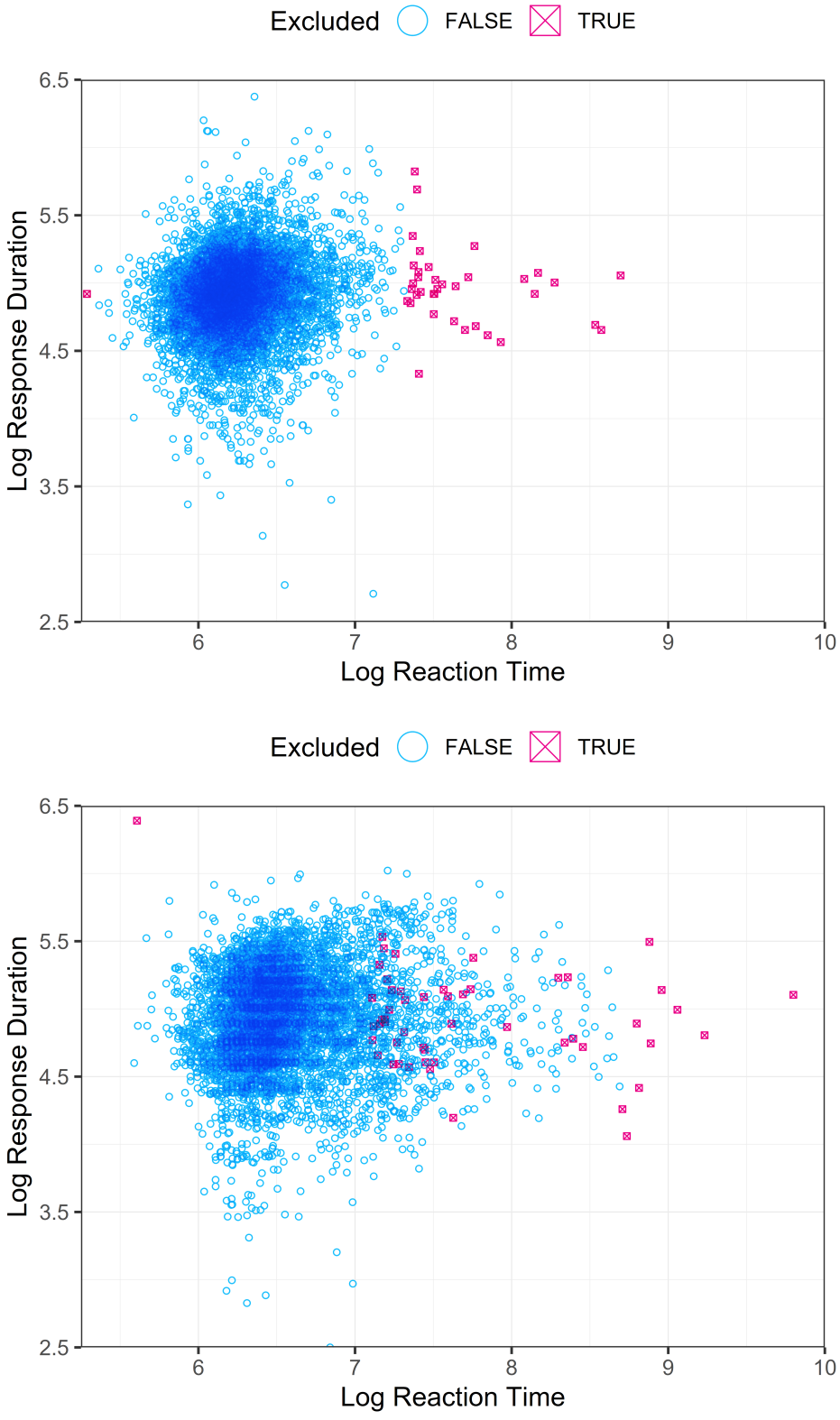


Figure C13: RT distributions for the in-person (Exp. 2, top) and the web-based (Exp. 2b, bottom) lexical decision experiments. The plots are zoomed in slightly for comparison; the top panel excludes a further 6 points beyond the bounds of the plot.

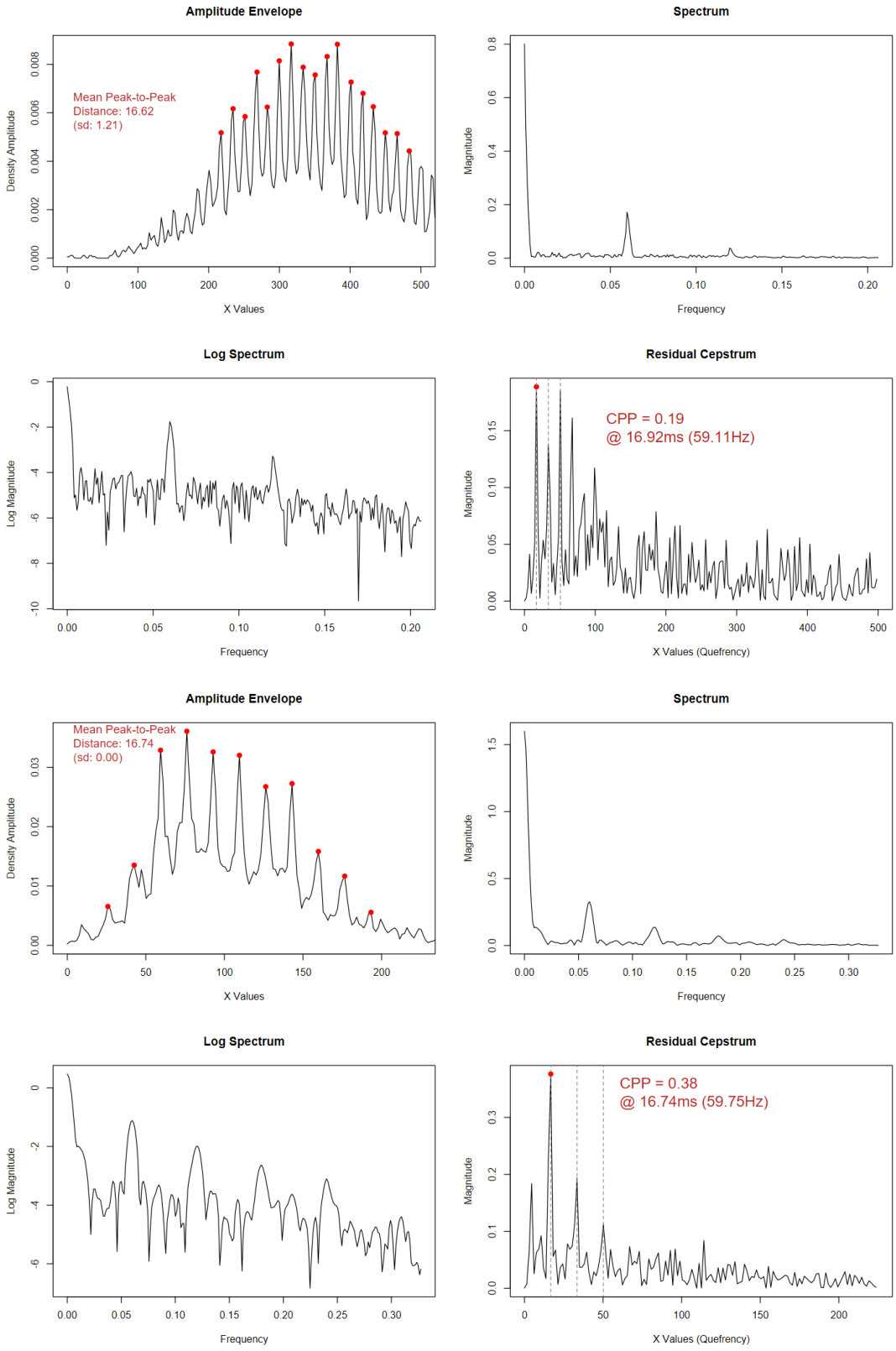


Figure C14: Cepstral analysis for web-based RT (top four) and RD data (bottom four). The cepstra in the bottom right panels are residualized, i.e., subtracting out the overall exponential decay trend. The cepstra show clear harmonic structure (harmonics 1-3 shown with dashed lines) expected from robust periodicity (c.f. lack of a clear pattern in Figure C15 (p. 262)).

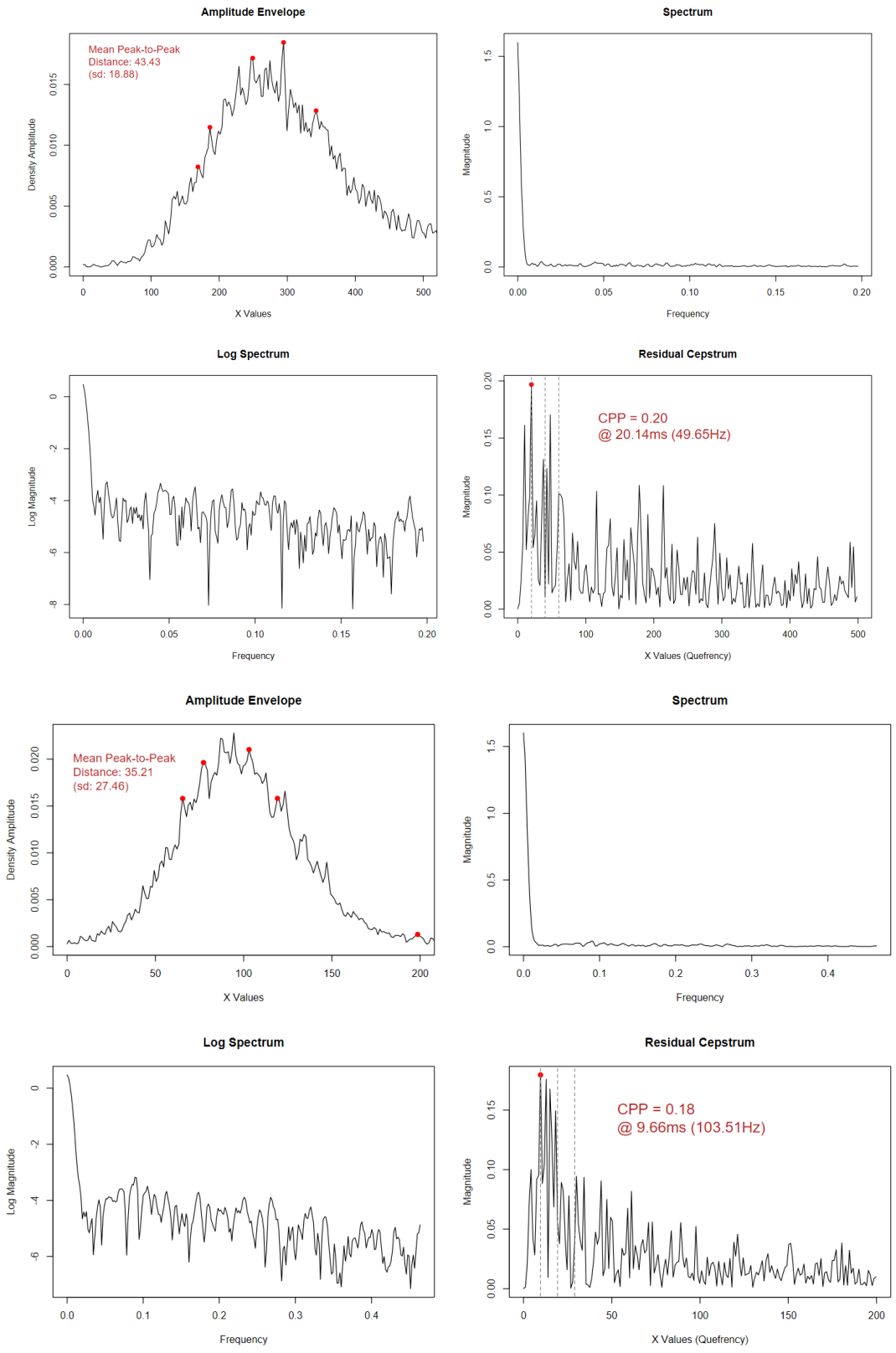


Figure C15: Cepstral analysis for the in-person experiment (Exp. 2) RT (top four) and RD data (bottom four). The cepstra do not show clear peaks at the harmonics, nor is the estimated CPP robust or related to the refresh rate of the monitor (c.f. clear patterns in Figure C14 (p. 261)).

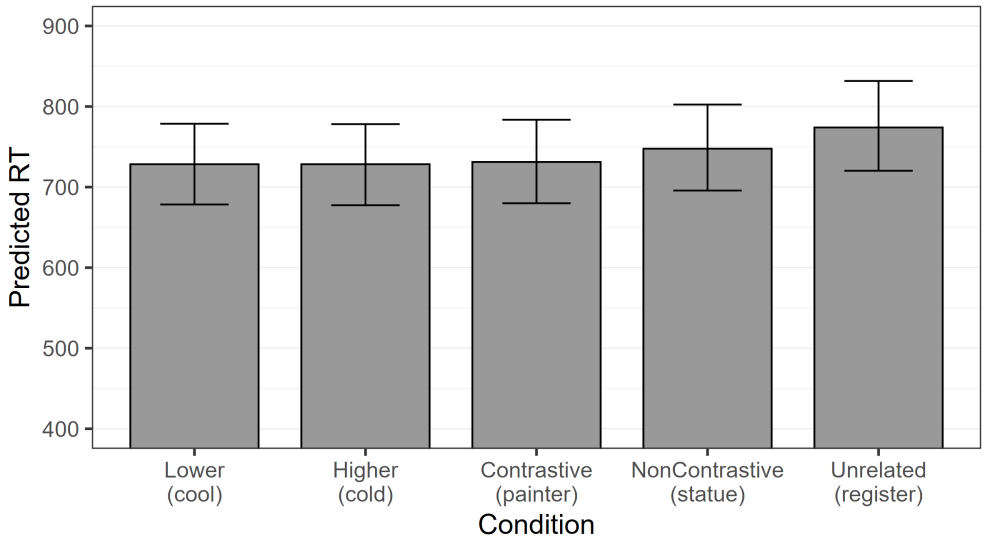


Figure C16: Web-based lexical decision (Exp. 2b) posterior predicted reaction times with 95% credible intervals for each target condition in the online experiment.

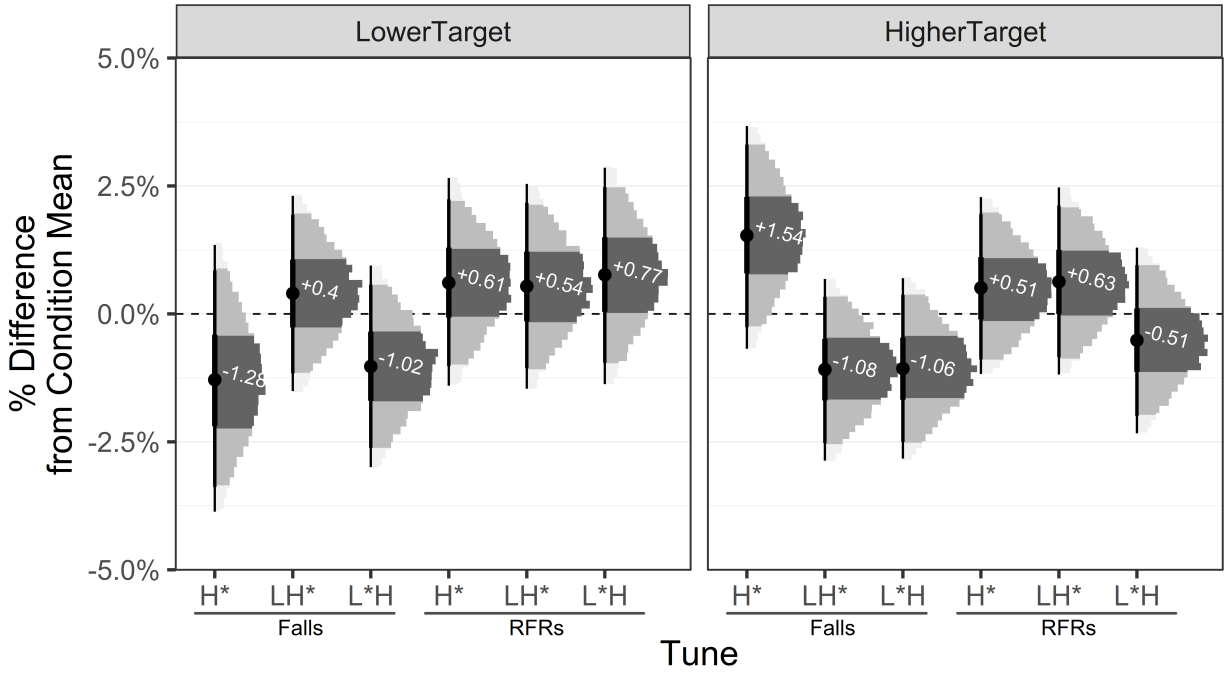


Figure C17: Web-based lexical decision (Exp. 2b), posterior predicted percent change distributions for each tune in the LOWERTARGET and HIGHERTARGET conditions.

C.6 Exp. 3 (In-Person 0ms SOA Lexical Decision)

The full model results are reported in Table C6 (p. 266).

C.7 Exp. 4 (Dual Task) Details

The by-tune results for the dual task are shown in Figure C18 (p. 264). As mentioned in the main text, generally the results are similar to the online experiment results, where the subtle effect of tune is not robust to the switch cost involved in the task.

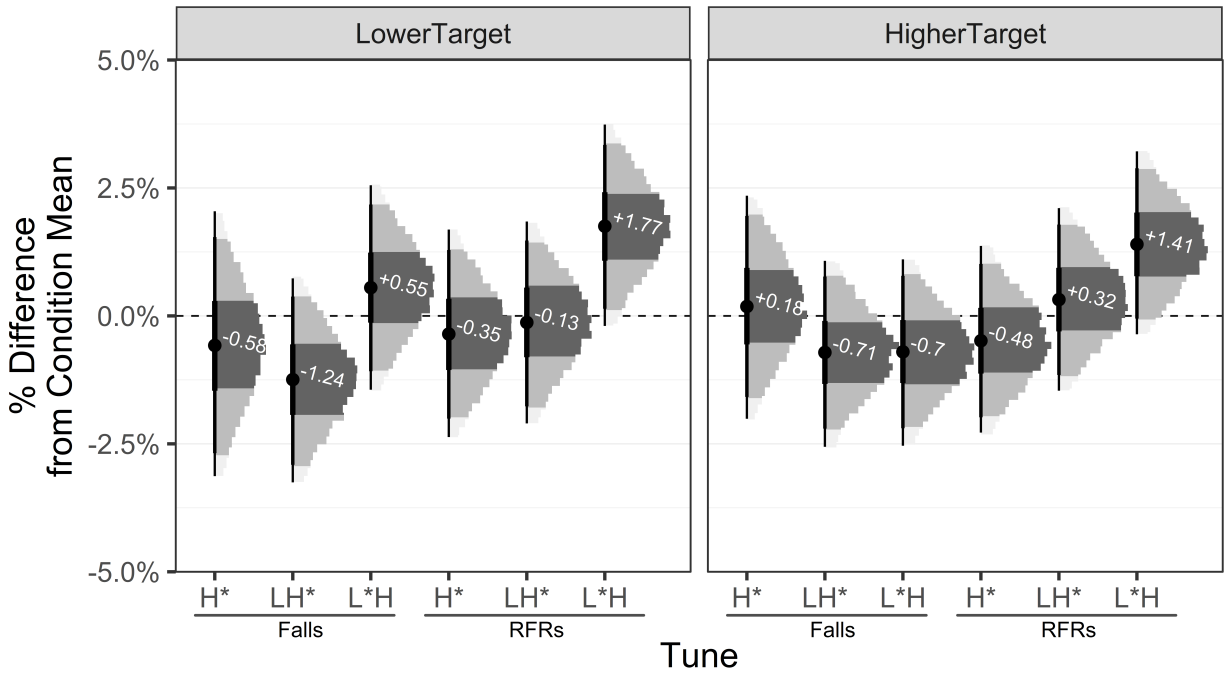


Figure C18: Exp. 4 (dual task) posterior mean percent change distributions with 50/89/95% CrI.

The model-predicted RTs for each condition, split by the participant's Yes/No response, are also shown in Figure C19 (p. 265).

Figure C20 (p. 268) shows the distribution of RTs and RDs for both components of the dual task; see Figure C13 (p. 260) for the RTs for the online and in-person lexical decision RT distributions.[6] Table C7 (p. 269) shows the full model results for the lexical decision RT statistical

[6]A lognormal model was fit to test the difference in RTs for the standalone lexical decision task (Exp. 2) and the lexical decision portion of the dual task (Exp. 4). The model uses formulas of $\mu \sim$`experiment*condition +`

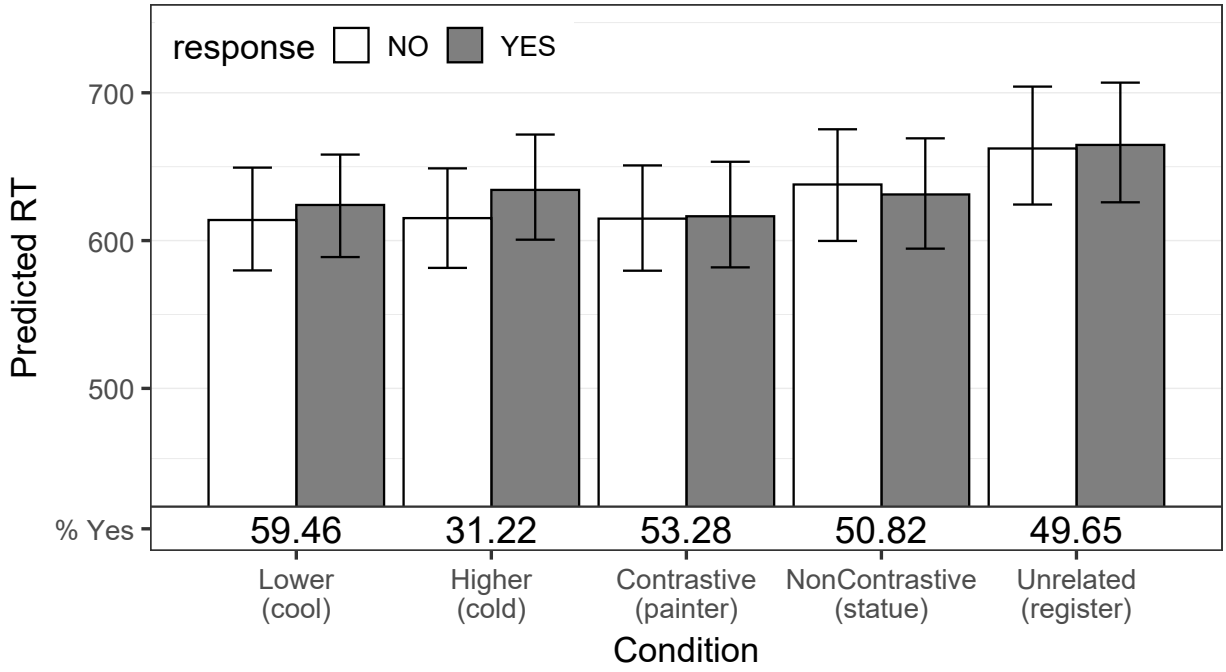


Figure C19: Exp. 4 (dual task) posterior-predicted RTs with 95% CrI for each target condition. Example questions were *Would you conclude that the office was not merely cool?* (LowerTarget); *Would you conclude that the office was not cold?* (HigherTarget); *Would you conclude that the museum did not thrill the painter?* (Contrastive/NonContrastive/Unrelated). Values below bars show the empirical proportion of Yes responses for each condition (e.g., overall SI rate is 31.22%).

Term	Estimate	%Δ	SE	95% CrI	PD
LowerTarget Mean	6.252		0.020	[6.214, 6.290]	100.00
WordLogFrequency	−0.036	−3.58	0.003	[−0.043, −0.030]	100.00
WordLength	0.007	0.71	0.003	[0.002, 0.012]	99.48
Block	−0.021	−2.06	0.001	[−0.023, −0.018]	100.00
H*LH	−0.009	−0.87	0.008	[−0.025, 0.008]	85.45
LH*LH	0.010	0.98	0.008	[−0.006, 0.026]	88.04
LH*LL	0.012	1.16	0.008	[−0.004, 0.028]	92.14
L*HLH	−0.019	−1.86	0.008	[−0.035, −0.003]	99.10
L*HLL	−0.013	−1.26	0.008	[−0.029, 0.004]	93.76
Higher−Lower	−0.002	−0.17	0.012	[−0.025, 0.022]	55.78
NonScalar−Scalar	−0.007	−0.70	0.013	[−0.033, 0.018]	70.47
Ncont−Contrastive	0.035	3.51	0.013	[0.009, 0.060]	99.55
Unrelated−Related	0.087	9.05	0.015	[0.057, 0.115]	100.00
Higher−Lower					
H*LH	0.008	0.75	0.012	[−0.015, 0.031]	74.25
LH*LH	−0.010	−0.98	0.011	[−0.032, 0.012]	80.55
LH*LL	−0.011	−1.12	0.012	[−0.034, 0.011]	83.01
L*HLH	0.028	2.83	0.011	[0.006, 0.050]	99.31
L*HLL	0.004	0.36	0.012	[−0.019, 0.026]	62.05
NonScalar−Scalar					
H*LH	−0.012	−1.20	0.020	[−0.051, 0.027]	73.71
LH*LH	0.022	2.24	0.019	[−0.015, 0.060]	87.92
LH*LL	−0.013	−1.30	0.019	[−0.050, 0.024]	76.04
L*HLH	0.007	0.74	0.020	[−0.030, 0.046]	64.78
L*HLL	−0.011	−1.14	0.019	[−0.049, 0.027]	72.49
Ncont−Contrast					
H*LH	0.024	2.44	0.020	[−0.016, 0.063]	88.56
LH*LH	−0.002	−0.17	0.022	[−0.045, 0.041]	53.38
LH*LL	−0.004	−0.45	0.019	[−0.041, 0.032]	59.59
L*HLH	0.003	0.34	0.021	[−0.037, 0.044]	56.60
L*HLL	−0.013	−1.29	0.019	[−0.050, 0.025]	75.28
Unrelated−Related					
H*LH	0.014	1.39	0.021	[−0.028, 0.055]	73.76
LH*LH	−0.014	−1.42	0.020	[−0.054, 0.026]	76.12
LH*LL	−0.008	−0.79	0.021	[−0.049, 0.033]	64.68
L*HLH	0.005	0.45	0.022	[−0.039, 0.047]	58.30
L*HLL	0.015	1.55	0.021	[−0.027, 0.057]	76.88

Table C6: Exp. 3 (short SOA) full statistical model summary.

`(1+condition|participant)` and σ ∼`experiment + (1|participant`—the latter models the variance parameter of the lognormal distribution. Experiment is a two-level categorical variable that is scaled sum coded with the dual task as the reference level. The model finds that lexical decision RTs in the standalone task task are indeed faster ($\hat{\beta} = -0.07, CrI : [-0.12, -0.03]$) and less dispersed ($\hat{\beta} = -0.21, CrI : [-0.29, -0.12]$) that RTs in the dual task.

models. Figures C21 (p. 270) and C22 (p. 271) show the by-item variation in SI rates. For the reader interested in the MI rates, Figures C23 (p. 272) and C24 (p. 273) show the by-item variation in the MI rates.

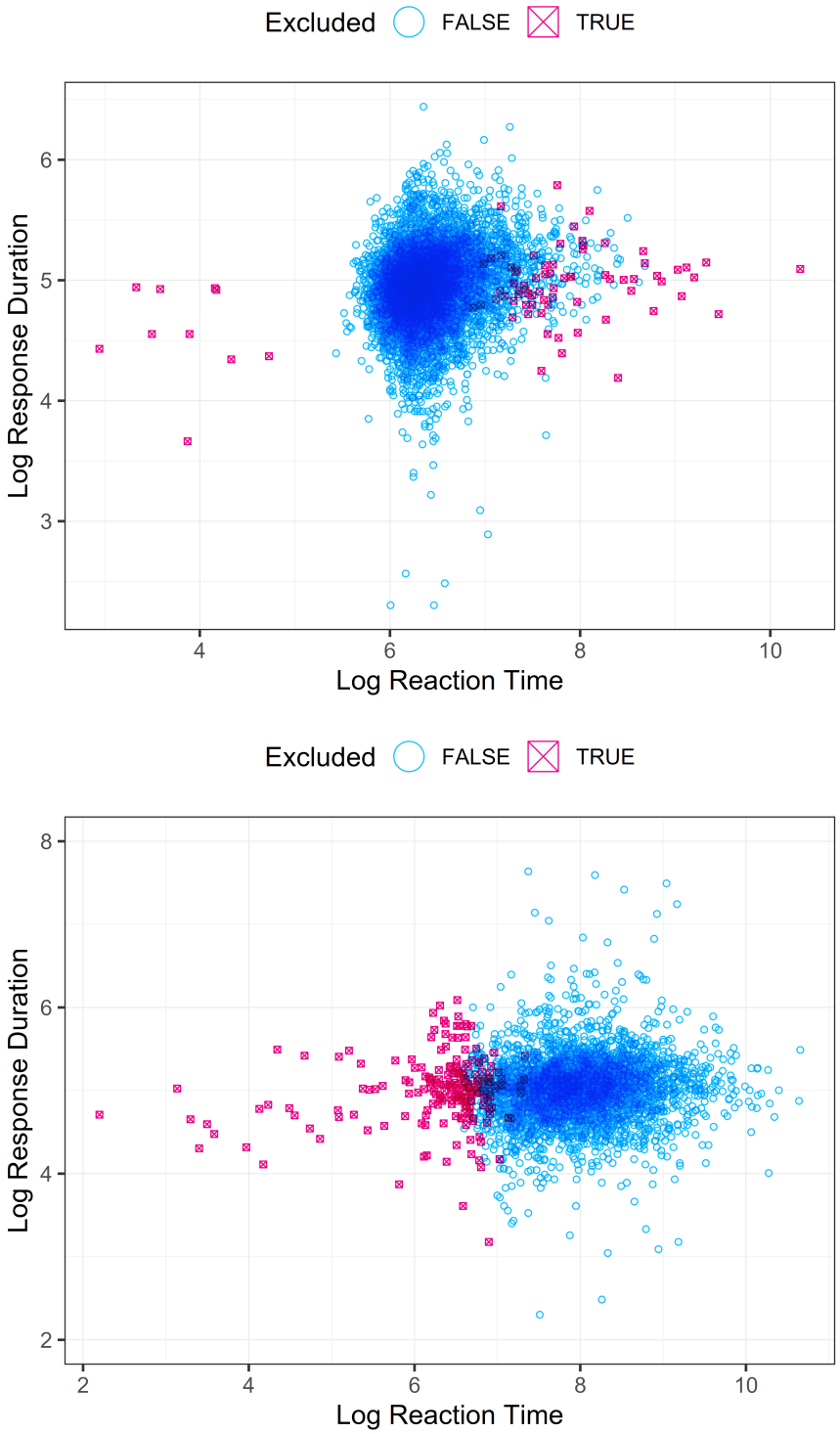


Figure C20: Dual task (Exp. 4) RT and RD distributions for the lexical decision portion (top) and the inference task portion (bottom).

Term	Estimate	%Δ	SE	95% CrI		PD
LowerTarget Mean	6.409		0.028	[6.352,	6.465]	100.00
WordLogFrequency	−0.035	−3.42	0.003	[−0.040,	−0.029]	100.00
WordLength	0.007	0.72	0.003	[0.002,	0.012]	99.75
Block	−0.029	−2.81	0.001	[−0.031,	−0.026]	100.00
H*LH	−0.005	−0.48	0.009	[−0.023,	0.014]	69.33
LH*LH	0.003	0.32	0.009	[−0.015,	0.021]	63.12
LH*LL	−0.007	−0.71	0.009	[−0.026,	0.011]	78.08
L*HLH	0.014	1.41	0.009	[−0.004,	0.032]	93.81
L*HLL	−0.007	−0.70	0.009	[−0.025,	0.011]	77.78
Higher−Lower	0.003	0.35	0.012	[−0.020,	0.027]	61.42
NonScalar−Scalar	−0.007	−0.69	0.013	[−0.032,	0.018]	70.79
Ncont−Contrastive	0.024	2.42	0.012	[0.000,	0.048]	97.65
Unrelated−Related	0.063	6.52	0.013	[0.038,	0.088]	100.00
Higher−Lower						
H*LH	−0.001	−0.13	0.013	[−0.026,	0.024]	54.20
LH*LH	0.004	0.45	0.013	[−0.020,	0.030]	63.06
LH*LL	0.005	0.53	0.013	[−0.020,	0.030]	65.60
L*HLH	−0.004	−0.35	0.013	[−0.029,	0.022]	60.92
L*HLL	−0.013	−1.25	0.013	[−0.038,	0.013]	83.08
NonScalar−Scalar						
H*LH	−0.014	−1.43	0.019	[−0.051,	0.024]	77.88
LH*LH	0.025	2.52	0.019	[−0.012,	0.061]	90.48
LH*LL	−0.011	−1.14	0.018	[−0.047,	0.024]	73.51
L*HLH	0.013	1.28	0.019	[−0.026,	0.050]	75.11
L*HLL	−0.018	−1.76	0.019	[−0.055,	0.020]	82.69
Ncont−Contrast						
H*LH	0.010	1.02	0.018	[−0.025,	0.044]	72.17
LH*LH	0.004	0.44	0.019	[−0.033,	0.041]	60.06
LH*LL	0.007	0.74	0.018	[−0.027,	0.042]	66.39
L*HLH	0.004	0.38	0.020	[−0.036,	0.043]	58.03
L*HLL	−0.010	−0.96	0.019	[−0.047,	0.027]	69.67
Unrelated−Related						
H*LH	0.021	2.07	0.020	[−0.019,	0.058]	85.51
LH*LH	−0.017	−1.67	0.019	[−0.054,	0.021]	80.38
LH*LL	0.007	0.69	0.021	[−0.034,	0.048]	63.39
L*HLH	−0.024	−2.40	0.020	[−0.065,	0.016]	87.97
L*HLL	0.011	1.10	0.021	[−0.031,	0.050]	70.91

Table C7: Exp. 4 (dual task) full statistical model summary.

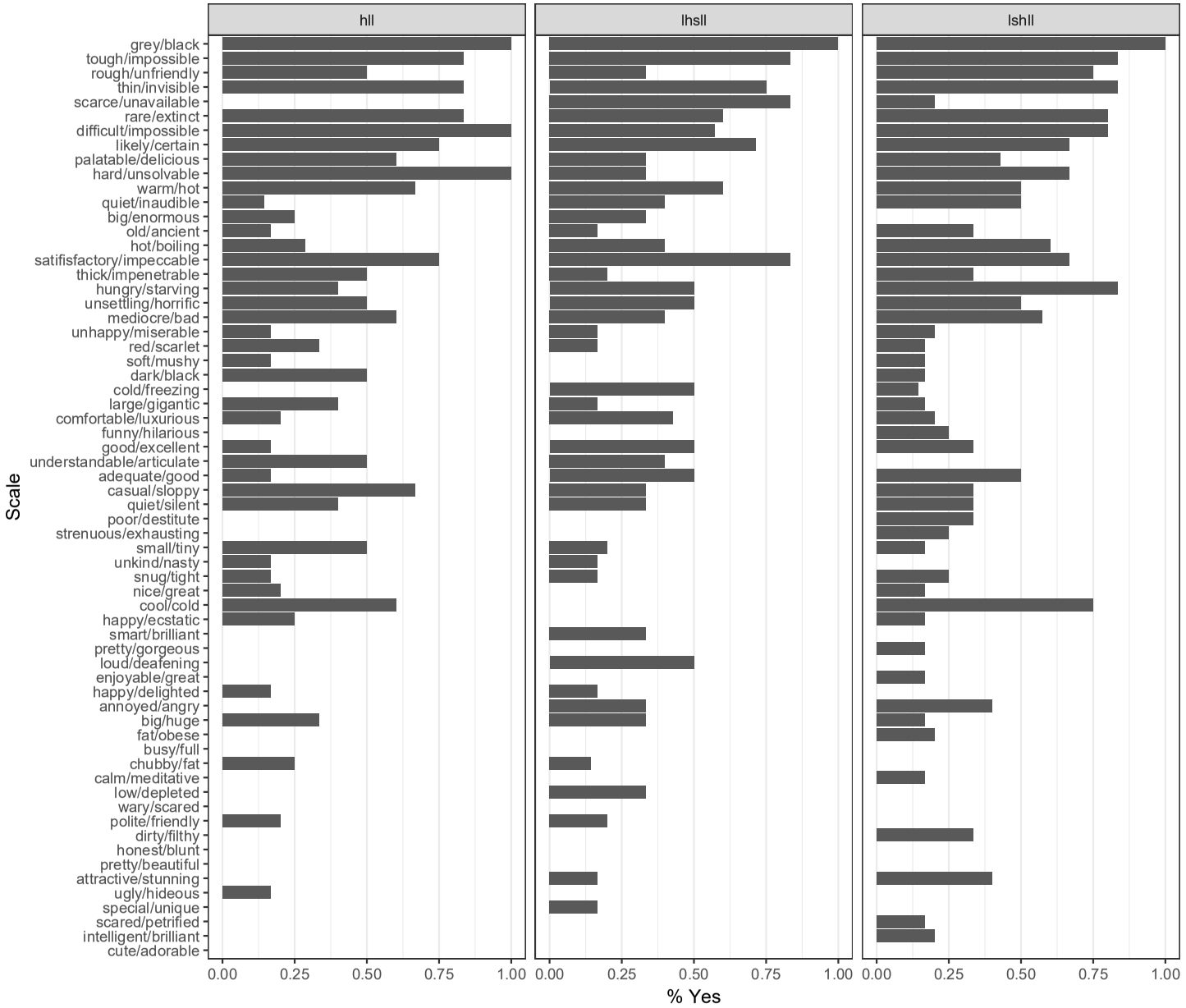


Figure C21: Dual task (Exp. 4) by-item SI rates for Falling tunes. Scales are ordered in the same way as Exp. 1.

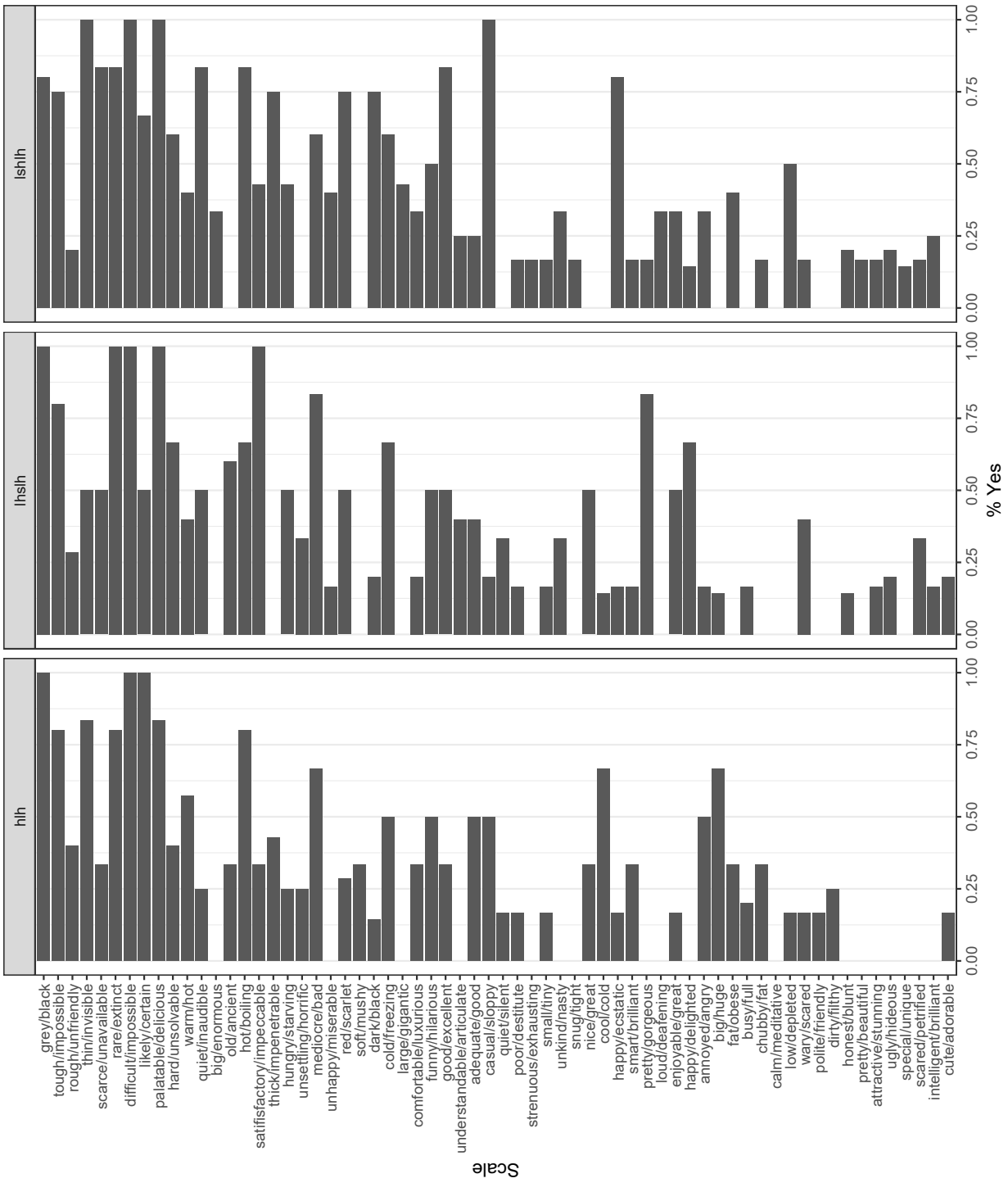


Figure C22: Dual task (Exp. 4) by-item SI rates for RFR-shaped tunes. Scales are ordered in the same way as Exp. 1.

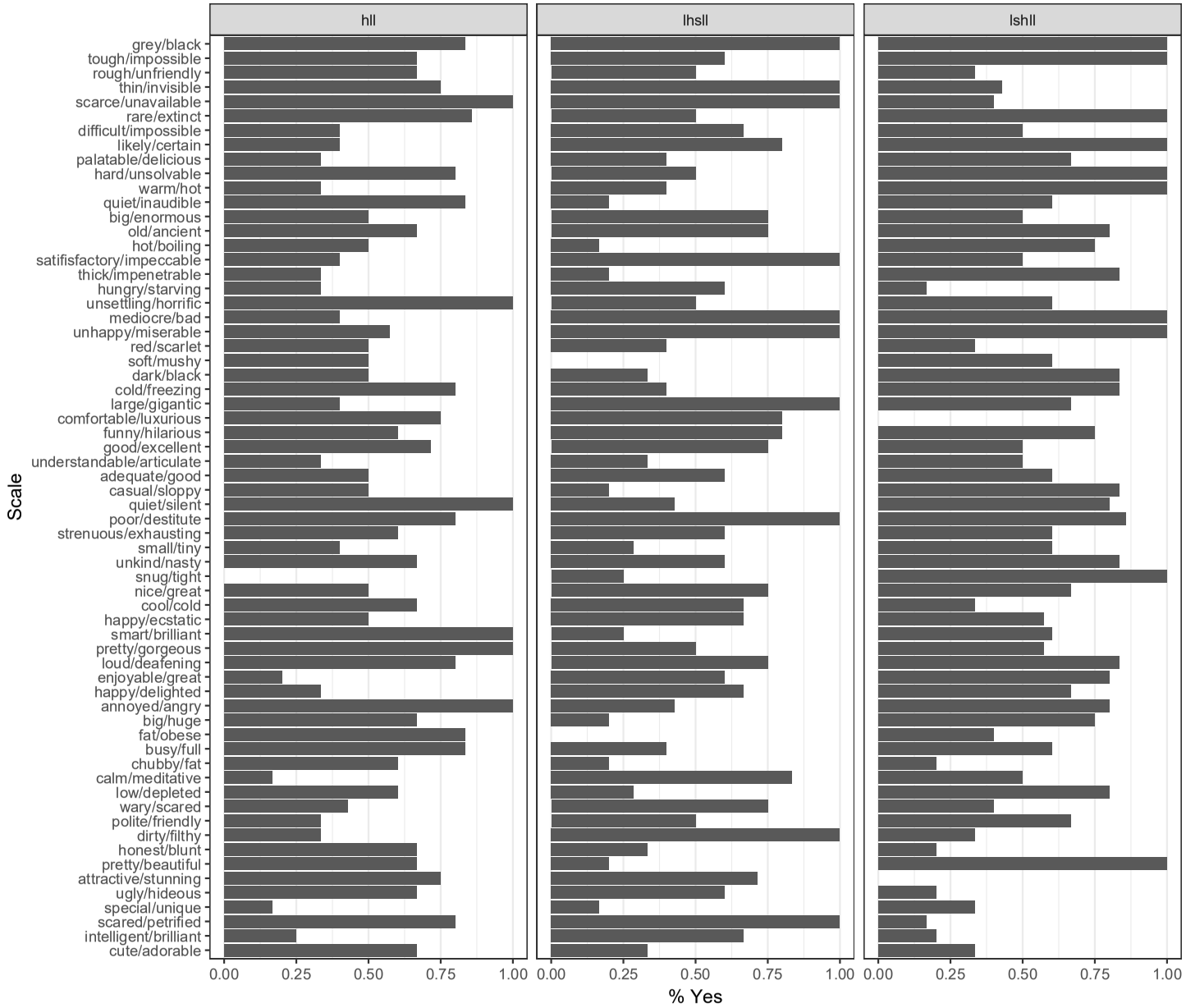


Figure C23: Dual task (Exp. 4) by-item MI rates for Falling tunes. The scales are ordered in the same way as Exp. 1.

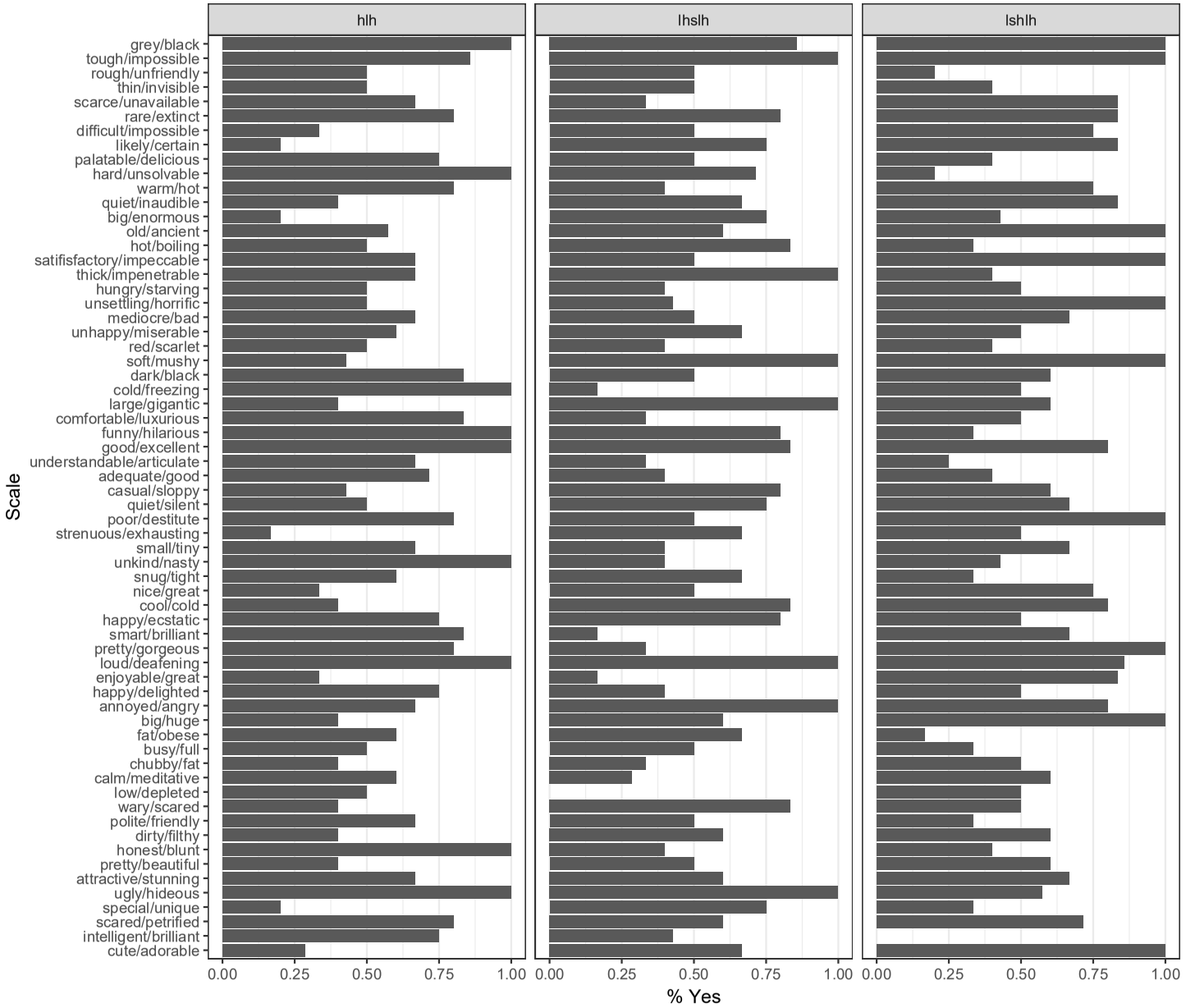


Figure C24: Dual task (Exp. 4) by-item MI rates for RFR-shaped tunes. Scales are ordered in the same way as Exp. 1.

STRUCTURED VARIATION IN INTONATIONAL FORM AND INTERPRETATION IN AMERICAN ENGLISH

Approved by:

Jennifer Cole
Linguistics
Northwestern University

Gregory Ward
Linguistics
Northwestern University

Eszter Ronai
Linguistics
Northwestern University

Date Approved: April 25, 2025