

Using intonation to disambiguate meaning:

The role of empathy and proficiency in L2 perceptual development

Joseph V. Casillas, Juan José Garrido Pozú, Kyle Parrish, Laura Fernández Arroyo, Nicole Rodríguez, Robert Esposito, Isabelle Chang, Kimberly Gómez, Gabriela Constantin-Dureci, Jiawei Shao, Iván Andreu Rascón, & Katherine Taveras

Rutgers University

Author note

The authors made the following contributions. Joseph V. Casillas: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing - Original Draft Preparation, Writing - Review & Editing; Juan José Garrido Pozú: Conceptualization, Investigation, Methodology, Resources, Writing - original draft, Writing - Review & Editing; Kyle Parrish: Conceptualization, Investigation, Methodology, Writing - original draft, Writing - Review & Editing; Laura Fernández Arroyo: Conceptualization, Investigation, Methodology, Writing - original draft, Writing - Review & Editing; Nicole Rodríguez: Conceptualization, Investigation, Methodology, Writing - original draft, Writing - Review & Editing; Robert Esposito: Conceptualization, Investigation, Methodology, Writing - Review & Editing; Isabelle Chang: Conceptualization, Investigation, Methodology, Writing - original draft, Writing - Review & Editing; Kimberly Gómez: Conceptualization, Investigation, Methodology, Writing - original draft; Gabriela Constantin-Dureci: Writing - Review & Editing; Jiawei Shao: Writing - original draft, Writing - Review & Editing; Iván Andreu Rascón: Writing - original draft, Writing - Review & Editing; Katherine Taveras: Writing - Review & Editing.

Correspondence concerning this article should be addressed to Joseph V. Casillas, Rutgers University - Department of Spanish and Portuguese, 15 Seminary Place, New Brunswick, NJ 08904, USA . E-mail: joseph.casillas@rutgers.edu

Abstract

The present study investigates the interplay between proficiency and empathy in the development of second language (L2) prosody by analyzing the perception and processing of intonation in questions and statements in L2 Spanish. A total of 225 adult L2 Spanish learners (L1 English) from the Northeastern United States completed a two-alternative forced choice (2AFC) task in which they listened to four utterance types and categorized them as either questions or statements. We used Bayesian multilevel regression and Drift Diffusion modeling to analyze the 2AFC data as a function of proficiency level and empathy scores for each utterance type. We show that learner response accuracy and sensitivity to intonation is positively correlated with proficiency, and this association is affected by individual empathy levels in both response accuracy and sentence processing. Higher empathic individuals, in comparison with lower empathic individuals, appear to be more sensitive to intonation cues in the process of forming sound-meaning associations, though increased sensitivity does not necessarily imply increased processing speed. The results motivate the inclusion of measures of pragmatic skill, such as empathy, to better account for intonational meaning processing and sentence comprehension in second language acquisition.

Keywords: Second language acquisition, Sentence processing, Speech perception, Intonation, Empathy

Word count: 9,875

Using intonation to disambiguate meaning:

The role of empathy and proficiency in L2 perceptual development

A fundamental difficulty of speech comprehension is that listeners can come to understand different messages when presented with the same linguistic information (Cain, Oakhill, & Lemmon, 2004). This can be especially problematic when one begins the endeavor of learning a new language. In particular, it is common for second language (L2) learners to struggle with intonation—i.e., the melodic contour of an utterance—in the target language (Trofimovich & Baker, 2006). The difficulties associated with intonation can result in comprehension and communication mishaps because the tune is associated with numerous parts of the linguistic system, such as sentence function, e.g., utterance type, syntactic constituency, as well as pragmatic function, e.g., information structure (Casielles-Suárez, 2004; Erteschik-Shir, 2007), speaker belief states (Pierrehumbert & Hirschberg, 1990), polite discourse (Astruc, Vanrell, & Prieto, 2016), bias, or presupposition (Henriksen, Armstrong, & García-Amaya, 2016). The present study investigated how the perception of intonation during sentence processing develops in adult L2 learners.

Recent research on monolingual populations suggests that individual differences in pragmatic skills, such as empathy, may play a role in meaning disambiguation (Aziz-Zadeh, Sheng, & Gheytanchi, 2010; Bishop, 2016; Esteve-Gibert, Portes, Schafer, Hemforth, & D'Imperio, 2016; Esteve-Gibert et al., 2020; Orrico & D'Imperio, 2020). Concretely, higher empathy individuals, in comparison with lower empathy individuals, appear to be more sensitive to the intonational cues of speech during the process of forming sound-meaning associations. Furthermore, increased attention has been given to how individual differences in learner backgrounds play a role in the process of L2 acquisition (Hu et al., 2013; Liu, 2017; Rota &

Reiterer, 2009). The present study contributes to these lines of research by examining how individual differences in pragmatic skills affect the development of intonation during sentence comprehension. Specifically, we investigated the interplay between language proficiency and an individual pragmatic skill (empathy) when learning an L2. We focused on the role of empathy in the development of L2 prosody by analyzing the perception of intonation in questions and statements in L2 Spanish. In addition, we considered the role of dialectal variation by exposing listeners to utterances from eight varieties of Spanish.

Background and motivation

L2 acquisition of prosody. The difficulties associated with learning an additional language in adulthood are numerous. More often than not the focus falls on individual sounds, or segments, though there is evidence that adults who learn an L2 face suprasegmental challenges as well (Craft, 2015; Thornberry, 2014; Trofimovich & Baker, 2006, among others). Concretely, L2 learners often struggle with intonation, i.e., melodic variation at the utterance level. This is, in part, because in everyday discourse speakers can use intonation for numerous communicative functions, such as indicating syntactic structure, signaling pragmatic meaning, e.g., whether an utterance is a question or a statement, focusing constituents, conveying affective meaning, etc. Notably, the manner in which intonation is mapped to meaning is language-specific. As a consequence, L2 intonation is often produced in a non-target-like fashion due to cross-linguistic influence.

Intonation has a semantic function and through adequate cognitive decoding of the signal a listener can interpret the intended meaning of a given utterance. For example, an intonational contour can indicate to a listener whether the utterance of an interlocutor is a question or a statement. As touched upon above, a speaker can use prosody to signal numerous additional

pragmatic functions as well. For example, an information-seeking yes-no question can be contrasted with an echo yes-no question in Chilean Spanish by using L+H* HH% or L* HH% nuclear contour, respectively (See Ortiz-Lira, 2003; Ortiz-Lira & Cid-Uribe, 2000). This rich variation in pragmatic uses makes the interpretation and decoding of intonational contours during speech comprehension a non-trivial task for the language learner. Moreover, the use of first language (L1) prosodic features when speaking the target language can result in misunderstandings because the same prosodic features can convey different linguistic and paralinguistic meaning in the target language (Chen, 2005; Cruz-Ferreira, 1987; Mennen, 2007; Pickering, 2001). As noted by Levis (2016), prosody is also “[...] critical for L2 pronunciation because it plays a major role in cementing social bonds as a key marker of social identity” (p. 154).

For learners interested in obtaining native-like pronunciation, intonation is particularly relevant, as prosodic features have been found to be important cues in the perception of non-target-like accents, above and beyond other features of language (Jilka, 2000; Munro, 1995; Pettorino, De Meo, & Vitale, 2014). Nonetheless, intonation is not traditionally taught in the L2 classroom, perhaps because it is not common knowledge that proper control of prosody allows the learner not only to produce speech that is more intelligible, but also to comprehend speech in varied communicative settings (de-la-Mota, 2019; Derwing & Munro, 2015). The primary focus is generally placed on syntax and morphology, with target language phonology receiving much less, if any, attention (Rao, 2019). When target language pronunciation is addressed, it often focuses on segmental elements (de-la-Mota, 2019), despite the fact that merely being intelligible at the segmental level does not necessarily imply one will be pragmatically understood. As a result, some research has found that intonation is one of the last aspects of L2 phonology that learners acquire (e.g., Kvavik & Olsen, 1974).

Research on L2 intonation has been concerned primarily with speech production. Learner difficulties tend to be ascribed to L1 transfer, and models of L2 phonology, by and large, focus on the speech segment, as in the Speech Learning Model revised (Flege & Bohn, 2021), or contrasts between segments, i.e., PAM-L2, L2LP (Best & Tyler, 2007; Van Leussen & Escudero, 2015, respectively). Theoretical work centered on prosody in the acquisition of L2 phonology is relatively much less common, though some researchers have considered how the aforementioned models might account for suprasegmental phenomena (See Trofimovich & Baker, 2006). One clear example of this is the L2 Intonation Learning Theory (LILt, Mennen, 2015). LILt incorporates the basic assumptions of the SLM and PAM-L2, that L2 categories similar to L1 categories may be assimilated, but L2 categories that are perceptually different may be incorporated as new categories. Under this model, cross-language differences may occur along one or more intonation dimensions (systemic, realizational, semantic, and frequency) (See also Ladd, 2008) and the age of onset of acquisition may influence the degree of success in acquiring elements in different dimensions of language variation.

A dearth of knowledge remains regarding how perception of intonation develops in L2 learning, and even less is known about how individual pragmatic differences account for learner outcomes. Similar to the SLM, LILt focuses mostly on intonation production rather than perception, and adopts the assumption that difficulties in intonation production are perceptually motivated. The purpose of the present project is to address this gap in the literature by examining the perception of intonation during adult L2 phonological acquisition. For the present study, investigating L2 perception of intonation in statements and questions in L2 learners of Spanish provides an opportunity to examine how L2 perception develops and may differ from L1 perception, especially along the ‘semantic dimension’ of the LILt model, which focuses on how intonation is used to convey meaning. Importantly, whereas LILt considers the influence of

external factors such as age of acquisition on the success of learners, the present study investigates the role of empathy as a pragmatic skill on L2 acquisition of intonation, which contributes to our understanding of intonation development along a different dimension.

Acquisition of Spanish prosody. As with all phonetic phenomena, a lack of invariance in the acoustic content of prosodic realizations also increases the difficulty of the learner's task. Beyond the level of the individual, however, dialectal differences can account for additional difficulties. Spanish is extensively spoken across the world, with relatively small geolocal differences between varieties when compared with other languages, such that speakers from distinct regions can still generally understand each other. That being said, phonetic variation is abundant. For instance, the pitch accent of the same utterance type—e.g., a broad focus statement—may be realized differently with regard to pitch movement and/or syllable duration depending on the variety. Intonational strategies can be different altogether. Consider information-seeking yes/no questions, which, in some varieties like Puerto Rican, Argentine, and Dominican Spanish, can be produced with a falling F0 contour (See, Armstrong, 2010; Gabriel et al., 2010; Willis, 2010, respectively). These examples illustrate between-variety variability because they can differ from the more common final rise found in many other varieties of Spanish (See Hualde & Prieto, 2015).

Previous research on the acquisition of Spanish prosody has primarily focused on the production of statements and questions, particularly in the study abroad context, using pre-, post-test designs (See Craft, 2015; Henriksen, Geeslin, & Willis, 2010; Thornberry, 2014; Trimble, 2013a, among others). Though the degree of improvement is variable based on a myriad of factors—such as context formality (Trimble, 2013a), use of Spanish (Henriksen et al., 2010; Trimble, 2013a), social integration (Trimble, 2013a), or the development of meaningful social

relationships with native speakers (Thornberry, 2014)—this line of research suggests that learners gradually acquire target-like intonation as they gain experience in the L2.

There is a paucity of research on the perception of Spanish intonation, but limited studies corroborate the general finding in speech production that mastery is indeed possible for adult learners (Brandl, González, & Bustin, 2020; Marasco, 2020; Nibert, 2005, 2006; Shang, 2022; Trimble, 2013b). For instance, Trimble (2013b) examined the perception of intonational cues in statements and yes/no questions in L1 English L2 Spanish adult learners that had studied abroad in Venezuela, Spain, or not at all. Using a gating task, Trimble (2013b) found that intonational cues that were absent from participants' L1 were difficult to perceive, though learners were more accurate with statements than questions, and that familiarity with the target variety improved accuracy. The investigation lends support to the general notion that the L2 intonation system develops in tandem with proficiency in Spanish, which was positively correlated with time spent studying abroad.

In a similar vein, Brandl et al. (2020) also investigated the perceptual development of intonation in questions and statements in L2 Spanish. Specifically, Brandl et al. (2020) examined the effect of L2 proficiency on the perception of broad-focus and narrow-focus statements and wh- and yes/no questions in adult L1 English L2 Spanish learners. The learners completed a forced-choice task in which they were presented audio and visual stimuli in matched and mismatched conditions. The participants' task was to determine whether the sentence presented aurally was the same as the sentence presented visually. Brandl et al. (2020) found that

perception and processing¹ of L2 intonation improved in conjunction with proficiency in Spanish, though it was conditional on the utterance type, with yes/no questions being more difficult to process and acquire when compared with statements. The authors concluded that perception of L2 intonation develops gradually in conjunction with L2 proficiency.

To summarize, the extant literature suggests that mastery of L2 perception of intonation seems feasible for adult learners, as processing speed and accuracy both improve as L2 proficiency increases. That being said, some utterance types present more difficulties than others. Furthermore, familiarity with the L2 variety can positively impact learner outcomes, which is particularly relevant given the rich phonetic and phonological variability attested in Spanish prosody. Much less is known regarding how perceptual development is modulated by individual differences, such as those related to pragmatic skill, though this is a recent and promising field of research that, moving forward, will help us understand individual variation (Bishop, Kuo, & Kim, 2020; See also Shang, 2022; Wiener & Bradley, 2020).

Empathy and pragmatic skill. The construct empathy refers to one's ability to infer the intentions of others. It is associated with understanding the feelings and emotions of those with whom one interacts (Baron-Cohen & Wheelwright, 2004). Research on empathy has associated the construct with Theory of Mind and perspective-taking (Baron-Cohen, 2011; Carruthers, 2009; Frith & Frith, 2003). Importantly, in recent years empathy has served as a proxy for investigating individual pragmatic skill. From the perspective of the listener, empathy is likely critical because it allows one to understand the intentions of others, predict their behavior, and understand their

¹ In the context of Brandl et al. (2020), ‘processing’ refers to *input processing* in adult second language acquisition (SLA), i.e., the strategies/mechanisms used by learners for linking linguistic form with meaning (See VanPatten, 2020).

emotions (Baron-Cohen & Wheelwright, 2004). Researchers that work on this construct have described two types of empathy that oftentimes might be difficult to distinguish. On the one hand, *affective empathy* represents one's ability to be emotionally aligned with the interlocutor and, on the other, *cognitive empathy* refers to recognizing and understanding the feelings and thoughts of an interlocutor. This suggests empathy is, to some degree, a necessary element when an individual seeks to understand and interact with its interlocutors in contexts involving literal and non-literal meaning.

The extant literature suggests that individual pragmatic skills modulate intonation processing (Bishop, 2016; Bishop, Chong, & Jun, 2015; Bishop & Kuo, 2016; Diehl, Bennetto, Watson, Gunlogson, & McDonough, 2008). Studies on monolingual populations show that individual pragmatic skills correlate with variability in semantic/pragmatic interpretation of ambiguous linguistic items (e.g., Degen & Tanenhaus, 2016; Nieuwland, Ditman, & Kuperberg, 2010). That is, in this line of research, individuals described as having higher pragmatic skill tended to prefer pragmatically enriched interpretations and individuals described as having less pragmatic skill tended to prefer more literal/semantic interpretations. In addition, more pragmatically skilled individuals have also been found to rely on different phonetic cues to parse syntactically ambiguous sentences when compared with less pragmatically skilled individuals (Bishop, 2016). Thus, one possibility is that variability in intonation perception is also linked to individual differences in pragmatic skills. A series of studies has investigated how empathy influences speech perception in monolingual populations (Esteve-Gibert et al., 2016, 2020; Orrico & D'Imperio, 2020). This work operationalizes the construct empathy as a pragmatic skill and has focused on it as a source of individual differences.

For instance, Esteve-Gibert et al. (2020) examined how listeners with different levels of empathy interpreted intonation and meaning in contexts in which a temporary lexical ambiguity could only be resolved through intonation. Empathy was measured using the Empathy Quotient (EQ, Baron-Cohen & Wheelwright, 2004), a self-report questionnaire, and participants were partitioned into groups corresponding with low or high empathy. Esteve-Gibert et al. (2020) tested French monolinguals in a visual-world paradigm eye-tracking task that resembled a card guessing game. Target objects were homophones in French (e.g., *cane*, Eng. “female duck”; *canne*, Eng. “walking stick”). Esteve-Gibert et al. (2020) found that processing of the lexical ambiguity (the homophones *cane/canne*) was modulated by empathy level when intonation was the only cue available. Specifically, highly empathic individuals varied their looking behavior as a function of intonational cues while less empathic individuals did not. That is, higher empathy individuals, in comparison with lower empathy individuals, were found to be more sensitive to intonation cues in the process of forming sound-meaning associations. In short, individuals with more pragmatic skill (higher empathy) appear to be able to use intonation to resolve temporary lexical ambiguities that can lead to confirmatory vs. contrasting interpretations. This research underscores the importance of considering individual pragmatic differences when examining intonational meaning processing and sentence comprehension.

Related research in the SLA context is scant, though early studies included affective variables—such as attitude, motivation, empathy, and, more recently, grit, among others—as they pertain to individual differences. Empirical studies on empathy are limited, though the construct received attention from scholars as early as the 60’s and 70’s (Brown, 1973; Guiora & Acton, 1979; Guiora, Beit-Hallahmi, Brannon, Dull, & Scovel, 1972; Guiora, Brannon, & Dull, 1972; See Guiora, Taylor, & Brandwin, 1968). The particular body of work linking empathy with SLA has focused on speech production, or, more specifically, on what early scholars considered

‘authentic pronunciation’ and, more recently, ‘pronunciation aptitude’ (See Rota & Reiterer, 2009), though no strong associations have been found. To the best of our knowledge, no studies have explored the construct empathy as it pertains to L2 perceptual development. Thus, at this time we do not know if empathy plays a role in L2 sentence processing in a similar manner to monolingual sentence processing. The present project extends this research to the SLA context to determine if individual differences in this pragmatic skill affect the development of intonation in L2 perception and sentence comprehension.

The present study

We investigate how proficiency and empathy are related to the development of L2 prosody by analyzing the perception of intonation in questions and statements in L2 Spanish. This study was pre-registered on the Open Science Framework² and designed to address the following research questions:

1. Is perceptual development in L2 Spanish modulated by proficiency and intonation type (i.e., Brandl et al., 2020)?
2. Do pragmatic skills—specifically, empathy—modulate the rate of development in L2 prosody?
3. Does speaker variety affect perception accuracy and processing speed?

Regarding RQ1, we hypothesize that accuracy will increase and processing time will decrease as a function of proficiency and intonation type. As shown in previous studies, yes/no questions (i.e., absolute interrogatives) will present the most difficulty for L2 learners of Spanish, followed by wh- questions (i.e., partial interrogatives) and broad focus and narrow focus statements. Based on the findings of Esteve-Gibert et al. (2020), we posit that prosodic development will occur

² See https://osf.io/dh4zp/?view_only=162d6d13e5814417bcb9de349f18cb62

sooner and at a faster rate in higher empathy individuals (RQ2). In this operationalization, ‘sooner’ refers to lower proficiency levels in a cross-sectional design, that is, at an earlier developmental stage when compared with lower empathy individuals. Finally, with regard to RQ3, we hypothesize that, overall, L2 learners will have the most difficulty (lower accuracy, slower response time) with the Cuban variety. This hypothesis is grounded in exploratory analyses of pilot data collected from 120 monolingual Spanish speakers in which responses to the Cuban variety were the least accurate (See Supplementary Materials for more information).

This project presents a conceptual replication of Brandl et al. (2020) in that we employ a similar experimental paradigm using similar stimuli in order to analyze the relationship between proficiency and L2 perception of intonation. Similar to Brandl et al. (2020), we include speakers from eight different varieties of Spanish in order to consider how dialectal variation influences perceptual development. We extend this research by taking into account pragmatic skill, specifically empathy, in L2 sentence processing. Importantly, this research builds on recent studies looking at the role of individual pragmatic skills in language comprehension and extends them to the field of SLA.

Method

Participants

Two hundred twenty-five individuals completed a two-alternative forced choice (2AFC) task in which auditory stimuli were identified as being questions or statements. Participants were recruited using the Prolific.ac online experimental platform and were compensated at a rate of \$9.52 per hour for their time. We estimated the task would take approximately 15 minutes to complete, thus each participant was paid \$2.70 for completing all three tasks. The mean time to

completion was approximately 13 minutes. The pool of participants was filtered using criteria set in Prolific.ac to ensure participants self-reported as being L1 English speakers born, raised, and currently living in the Northeastern US with no knowledge of any languages other than English or Spanish. They reported no hearing difficulties and were required to use headphones on a personal computer. Upon beginning the experiment, all participants responded to the following screening questions: 1) What part of the US are you from? 2) At what age did you begin learning Spanish? 3) Are you proficient in any languages other than English/Spanish? Additionally, participants responded to the prompt “I am most familiar with Spanish from...” and using a pull-down window they selected a variety of Spanish or “I am not familiar with any variety of Spanish”. We excluded data from any participant that responded that they were not from the US Northeast, that they began learning Spanish before the age of 13, or that they were proficient in a language other than English/Spanish. Participants responding categorically across all trials were also excluded. In sum, participants were adult native speakers of American English with varying levels of proficiency in Spanish, ranging from functionally monolingual to highly proficient. All participants with knowledge of Spanish were adult L2 learners, operationally defined as having begun the endeavor of learning Spanish after the age of 13.

Tasks

The study consisted of three tasks: a 2AFC task, a lexical decision vocabulary assessment, and a Likert-type questionnaire to assess empathy. The tasks were programmed in Python using Psychopy3 (Peirce et al., 2019) and presented online via Pavlovia. All code and materials used to generate the tasks are freely available on the Open Science Framework (https://osf.io/dh4zp/?view_only=162d6d13e5814417bcb9de349f18cb62).

2AFC. In the 2AFC task participants were presented an audio file containing a statement (broad focus or narrow focus) or a question (yes/no or wh-). Their task was to determine, as quickly and as accurately as possible, if the utterance they heard was a question or a statement. Specifically, they responded to an on-screen prompt asking “Is this a question?” using the keyboard. Participants typed ‘1’ for ‘yes’ (i.e., “yes, this is a question”) or ‘0’ for ‘no’ (i.e., “no, this is not a question”).

The auditory stimuli consisted of 64 critical items, 16 of each utterance type. The sentences were made up of three function words following SVO word order, which is the default in Spanish. The object was a noun with penultimate stress in all but three items. Subject pronouns were omitted in wh- questions. To generate the stimuli, we recorded native Spanish speakers of eight different varieties (Cuban, Peninsular-Madrileño, Peninsular-Andalusian, Puerto Rican, Chilean, Argentine, Mexican, and Peruvian). The eight native speakers all produced the same 64 critical items in a quiet room using professional recording equipment. The items were presented to the speaker on a screen. They were asked to read the item in silence to familiarize themselves with the context and to then read it aloud. To elicit narrow focus statements, one of the initiating authors read a question to the speaker and they responded. Table 1 provides an example of each utterance type.

Table 1: Example stimuli from the 2AFC task.

Utterance type	Prompt	Item	Translation
Broad focus statement	n/a	<i>Marta abre el regalo</i>	Marta opens the gift.
Narrow focus statement	¿Qué abre Marta?	<i>Marta abre el regalo</i>	(What does Marta open?) Marta opens the gift.
Wh- question	n/a	¿Por qué abre el regalo?	Why does she open the gift?
y/n question	n/a	¿Marta abre el regalo?	Does Marta open the gift?

All utterances were segmented using Praat (Boersma & Weenink, 2018) and normalized for peak intensity. A detailed description of the auditory stimuli is provided in the supplementary

materials. The 2AFC task included 64 trials in which the stimuli presented were randomized across speaker variety. Each variety had the same probability of being selected on a given trial, such that, on average, a given participant heard each variety approximately eight times (See online supplementary materials for more information). Prior to pre-registering our research questions and hypotheses, we piloted the 2AFC experiment on 120 monolingual Spanish speakers to assess the difficulty of the task and establish a baseline for response times. We did not come across any issues. An exploratory analysis of the monolingual data is provided in the supplementary materials.

LexTALE. To assess Spanish proficiency we administered the Lexical Test for Advanced Learners of Spanish (LexTALE-ESP, henceforth LexTALE) (Izura, Cuetos, & Brysbaert, 2014; Lemhöfer & Broersma, 2012). The LexTALE is a lexical decision experiment used to provide a standardized assessment of proficiency/vocabulary size in Spanish. In this task participants see a series of words on the computer screen and must decide if they are real or fake using the keyboard ('1' for real, '0' for fake). LexTALE scores can range from -20 to 60. Monolingual Spanish speakers generally score above 50. Scores from individuals with little or no knowledge of Spanish tend to be negative. Adult learners with low to medium proficiency can range from 0 to 25, and advanced learners generally score above 25. We conceive of proficiency as a continuous variable and therefore consider a monolingual English speaker to have little to no proficiency in Spanish (i.e., a negative value on the LexTALE). In our data set, participant scores ranged from -16 to 55, suggesting all proficiency levels were likely represented in the sample. The mean score was 12.95 (95% CrI: [11.18, 14.72]) with a standard deviation of 13.60 (95% CrI: [12.38, 14.9]).

Empathy Questionnaire. The construct empathy was assessed using the Empathy Quotient (EQ, Baron-Cohen & Wheelwright, 2004). The EQ is a 60-item questionnaire that presents four point Likert-type items ranging from ‘strongly agree’ to ‘strongly disagree’. Forty of the questions assess empathy and 20 are filler items. In order to avoid response bias, choices indicating empathic responses are coded to elicit “agree” responses in half the target items and “disagree” responses in the other half. The target items are scored with 2 or 1 points based on if the participant responds “strongly” or “slightly”. Finally, the EQ is scored by summing the total points to produce a single value indicating an individual’s level of empathy. Thus, the minimum possible value is 0 (low empathy) and the maximum is 80 (high empathy). In our data set the average empathy quotient was 37.88 (Range: [9, 69], 95% CrI: [36.13, 39.68], SD: 13.39, 95% CrI of SD: [12.28, 14.67]). The empathy quotient in its entirety is available in the supplementary materials.

Procedure

Participants recruited via Prolific.ac completed all three tasks in a single session. The 2AFC task was first, followed by the LexTALE task, and, finally, the empathy quotient questionnaire. We planned to collect data from approximately 300 individuals: 100 monolingual Spanish speakers not reported here and 200 L2 learners). Following Brandl et al. (2020), we assumed the effect size for perceptual learning was moderate in terms of the criteria set forth for L2 research by Plonsky and Oswald (2014) (Cohen’s D = 0.600, Pearson’s r = 0.287). Based on this assumption, we estimated that we would need 94 participants to have an 80% chance of capturing the proficiency effect with a type II error rate of 5%. Our hypothesis related to empathy as a possible mediator of intonation processing is exploratory in nature; therefore, we did not base our sample size estimate on any parameter estimates related to this effect. That said, we

believed the aforementioned exploratory effect was likely to be small, and, considering the resources necessary and available to us, planned to recruit 100 additional participants.

We excluded data from participants in the following circumstances: error during data collection, clear lack of understanding or engagement during the task (i.e., all ‘1’ responses, failed three attention checks, etc.), participants reporting having learned Spanish before the age of 13, or participants with knowledge of languages other than English and Spanish. Data from a total of 78 participants were discarded because the experimental session timed out and/or data was incomplete. An additional 8 participants were discarded due to low accuracy ($n = 5$), incomplete data ($n = 2$), and failed attention checks ($n = 1$). A total of 225 participants met the criteria for inclusion.

Statistical analyses

We report two primary statistical analyses that were pre-registered prior to collecting the learner data: response accuracy and drift diffusion models. All additional analyses are exploratory in nature and explicitly described as such. First, we analyzed response accuracy using Bayesian multilevel logistic regression. The model considered response accuracy for the population effects *utterance type* (broad focus statement, narrow focus statement, yes/no question, wh- question), *LexTALE score* (i.e., proficiency), *empathy quotient*, and the higher order interactions. The likelihood of the model was Bernoulli distributed with a logit link function. The criterion, *response*, was coded as ‘1’ for correct responses and ‘0’ for incorrect responses. Thus, the first analysis modeled the probability of responding correctly to the prompt “Is this a question?”. We specified group-level effects for participants, speaker variety, and items. The slope for *utterance type* varied for the participant effect, as did the *LexTALE* by *empathy quotient* interaction for the speaker variety effect. All continuous variables were standardized and ‘yes/no questions’ was set

as the baseline for *utterance type*, thus the model intercept represented the probability of a learner with average proficiency and average empathy responding correctly to a yes/no question.

The same model was fit to the response time data with the exception of the model likelihood, which was assumed to be distributed as lognormal. Response time was measured from the offset of the auditory stimuli. We arbitrarily excluded response times longer than ten seconds, which represented 37 tokens of 14,400 (0.26%). Participants were able to respond at any time after the onset of the auditory stimuli. There was a total of 443 (3.08%) tokens with negative response times. Of this subset, learners responded with 80.36% accuracy, therefore, we added the minimum value of the data set as a constant to all response times. As a result, the response time distribution comprised only positive values, a requirement of drift diffusion models (see below). We also fit an additional exploratory model with the same population- and grouping-effects structure using d' (d prime) as the outcome variable.

The second primary analysis utilized Bayesian drift diffusion modeling (DDM, Ratcliff & McKoon, 2008). This approach to analyzing behavioral data models decision-making as a random-walk decision process. DDMs can simultaneously take into account responses and response times in two-choice tasks in a single model, thus they are particularly beneficial when analyzing tasks in which speed-accuracy tradeoffs may be present. We estimate the parameters of the DDM using Bayesian methods and subsequently fit measurement error models on the posterior estimates of the resulting parameters.

A DDM estimates four parameters: boundary separation, bias, drift rate and non-decision time. Boundary separation, α , quantifies the amount of information necessary to make a decision. The boundaries represent the thresholds for the two alternatives in the task, which, in our case, implies correct and incorrect responses. Bias, β , gives an indication of a preference for one of the

choices at the beginning of the decision-making process. A positive bias value indicates a preference for the upper boundary, whereas a negative bias is an indicator of a preference for the lower boundary. The drift rate, δ , provides an assessment of the rate at which information is accumulated. A higher δ implies a random walk that arrives at one of the thresholds faster and is interpreted as an indication that the participant finds the task to be easier. Conversely, a lower drift rate is interpreted as indicating a more difficult task. The sign of the value is also relevant. Positive drift rate refers to evidence accumulation for the upper boundary and negative drift rate for the lower boundary. Finally, non-decision time, τ , models the part of the time course that is not associated with decision-making (e.g., the time necessary to perceive a stimulus prior to evidence accumulation). Figure 1 provides an example of a hypothetical DDM for the 2AFC task in the present project.

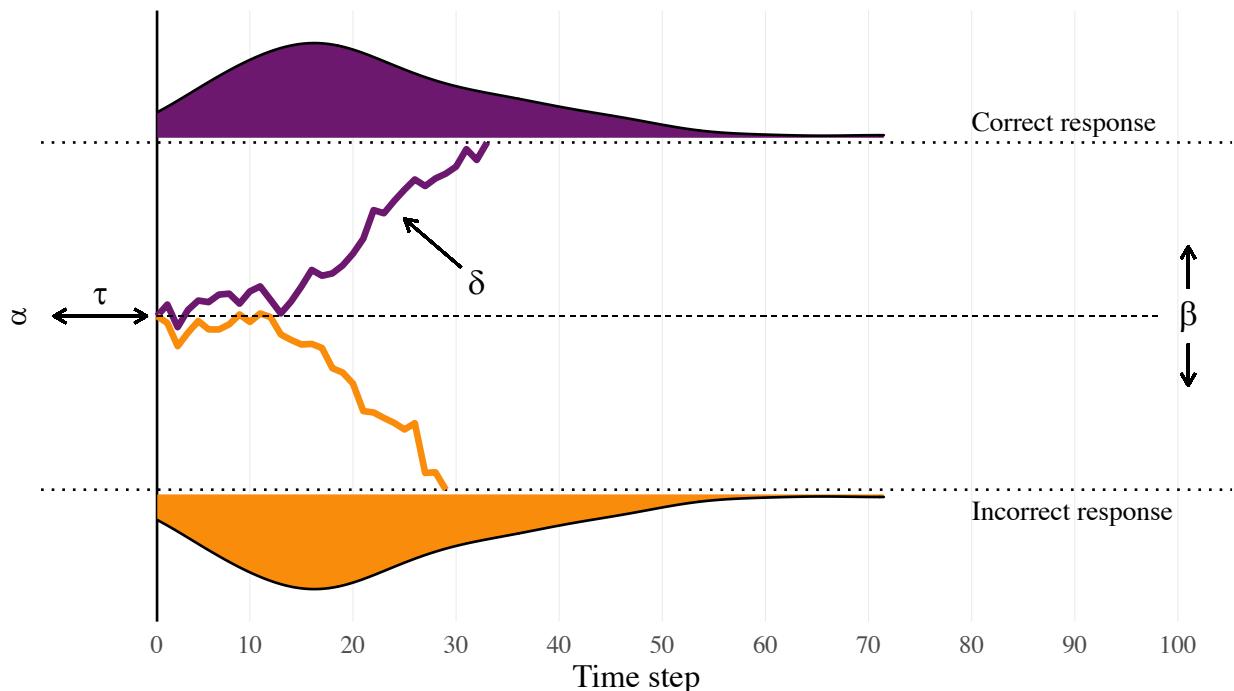


Figure 1. A drift diffusion model of the present study. The upper and lower bounds represent correct and incorrect responses, respectively. The boundary separation (ω) is the distance between the two thresholds and indicates the evidence required to make a decision. Non-decision time (τ) represents the time course before evidence accumulation begins, i.e., time used for any process except decision-making. Bias (β) is the starting point for the evidence accumulation in the

vertical plane (i.e., closer or further away from a given threshold), and drift rate (δ) quantifies the rate of evidence accumulation. The purple and orange lines represent examples of a decision resulting in a correct (purple) and incorrect (orange) decision. The corresponding density curves represent the distribution of response times at either threshold.

We estimated the aforementioned parameters by fitting a DDM to the response and response time data of each participant independently. We opted for this approach, as opposed to fitting a single model including all participants, for computational reasons. Put simply, the model likely would have taken weeks to fit, whereas the no-pooling (i.e., by-participant) method took approximately 26 hours. Thus, after fitting the DDMs, we obtained a posterior distribution of plausible values for boundary separation, drift rate, bias, and non-decision time for each participant. Next, we used measurement-error models to analyze boundary separation (α) and drift rate (δ) independently. These models followed the same functional form as the response accuracy model described above. That is, in two separate models, we analyzed the boundary separation and drift rate data as a function of *utterance type* (yes/no question, wh- question), *LexTALe score* (i.e., proficiency), *empathy quotient*, and the higher order interactions. The primary difference between the measurement-error models and the traditional regression analyses described for the response data is that the former can incorporate a measure of uncertainty around a point estimate. To give a concrete example, the analysis of the boundary separation data included the posterior median and the standard error for each participant as the outcome variable, as opposed to using just a single point estimate.

For all models, we included regularizing, weakly informative priors (Gelman, Simpson, & Betancourt, 2017). Generally, we sample from the posterior distribution of a given model for statistical inferences. To assess our pre-registered hypotheses we established a region of practical

equivalence (ROPE) around a point null value of 0 (see Kruschke, 2018) using the following formula:

$$\text{ROPE} = \frac{\mu_1 - \mu_2}{\sqrt{\frac{\sigma_1^2 + \sigma_2^2}{2}}}$$

For all models, median posterior point estimates are reported for each parameter of interest, along with the 95% highest density interval (HDI), the percent of the region of the HDI contained within the ROPE, and the maximum probability of effect (MPE). For statistical inferences, we focus on estimation rather than decision-making rules, though, generally, a posterior distribution for a parameter β in which 95% of the HDI falls outside the ROPE and a high MPE (i.e., values close to 1) are taken as compelling evidence for a given effect. All exploratory analyses, explicitly described as such, include posterior point estimates, the 95% HDI, and the MPE. We conducted all analyses using R and fit all models using the probabilistic programming language `stan` via the R package `brms` (Bürkner, 2017, 2018). Finally, we provide more information for all analyses in the supplementary materials.

Results

Response accuracy

Figure 2 (left panel) summarizes the posterior distribution of the omnibus response accuracy model, illustrating point estimates $\pm 66\%$ and 95% HDIs in graphical form. An equivalent summary of the posterior distribution in table format is provided in the supplementary materials (See Table 2). The log-odds of a correct response to a yes/no question at the average proficiency and EQ levels was 0.53, or approximately 62.95% ($\beta = 0.53$, HDI = [0.23, 0.82], ROPE = 0, MPE = 1). In comparison, all other utterance types were associated with an increase

in the log odds of responding correctly. The right panel of Figure 2 plots response accuracy of each utterance type in the probability space. As illustrated in the plot, participants were slightly more accurate when responding to wh- questions ($\beta = 0.43$, HDI = [0.17, 0.65], ROPE = 0, MPE = 1) with approximately 72.31% correct, and much more accurate when responding to declarative statements (narrow focus: $\beta = 2.13$, HDI = [1.84, 2.37], ROPE = 0, MPE = 1, Accuracy = 93.46%; broad focus: $\beta = 2.34$, HDI = [2.05, 2.59], ROPE = 0, MPE = 1, Accuracy = 94.63%).³

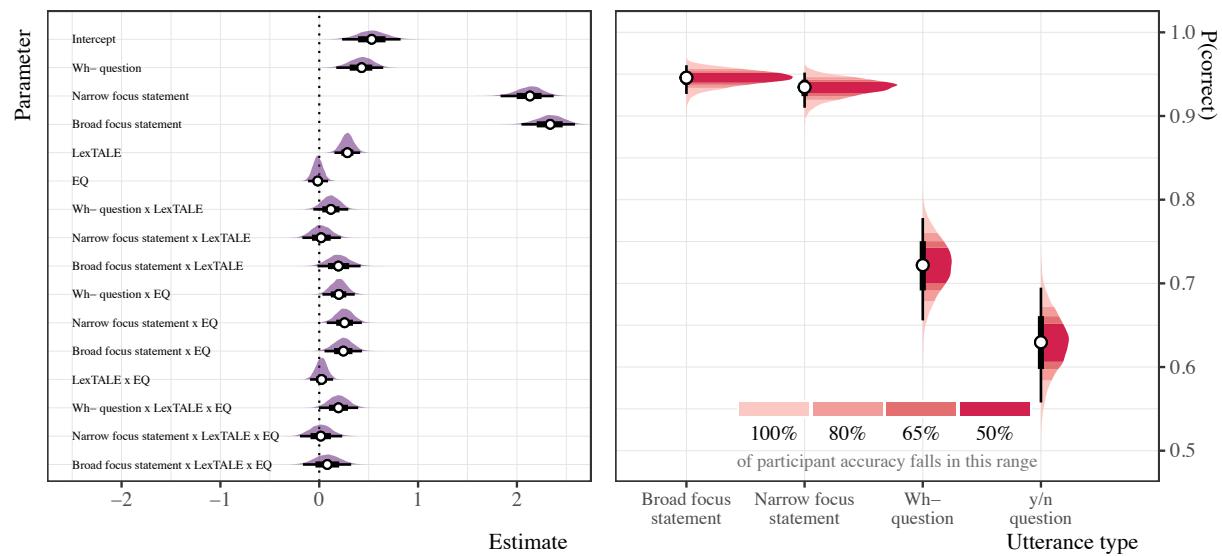


Figure 2. Forest plot summary of the response accuracy model (left panel) and posterior probability of a correct response for each utterance type (right panel). For both plots, white points represent posterior medians along with 66% and 95% highest density credible intervals.

Figure 3 plots response accuracy as a function of utterance type and proficiency (left panel) and empathy quotient (right panel). For all utterance types, response accuracy increased as proficiency increased. Though the proficiency effect was most visually obvious for yes/no questions ($\beta = 0.28$, HDI = [0.15, 0.41], ROPE = 0, MPE = 1) and wh- questions ($\beta = 0.40$, HDI

³ An exploratory (i.e., not pre-registered) analysis of sensitivity to utterance type was also conducted using d' in lieu of response accuracy. The results mirrored those found in the response accuracy model. That is, participants showed highest sensitivity to the declarative statements, followed by wh- and yes/no questions. These exploratory analyses are reported in the supplementary materials (See Table 4 and Figure 9).

$= [0.24, 0.57]$, ROPE = 0.00, MPE = 1.00), this was also the case for broad focus ($\beta = 0.48$, HDI = $[0.26, 0.71]$, ROPE = 0.00, MPE = 1.00) and narrow focus ($\beta = 0.31$, HDI = $[0.10, 0.51]$, ROPE = 0.00, MPE = 1.00) statements. There was no evidence that empathy level predicted response accuracy for yes/no questions ($\beta = -0.02$, HDI = $[-0.11, 0.09]$, ROPE = 0.98, MPE = 0.62), however for wh- questions ($\beta = 0.18$, HDI = $[0.05, 0.32]$, ROPE = 0.09, MPE = 1.00), broad focus statements ($\beta = 0.23$, HDI = $[0.04, 0.42]$, ROPE = 0.07, MPE = 0.99), and narrow focus ($\beta = 0.24$, HDI = $[0.07, 0.42]$, ROPE = 0.03, MPE = 1.00) statements we find compelling evidence that the effect is positive.

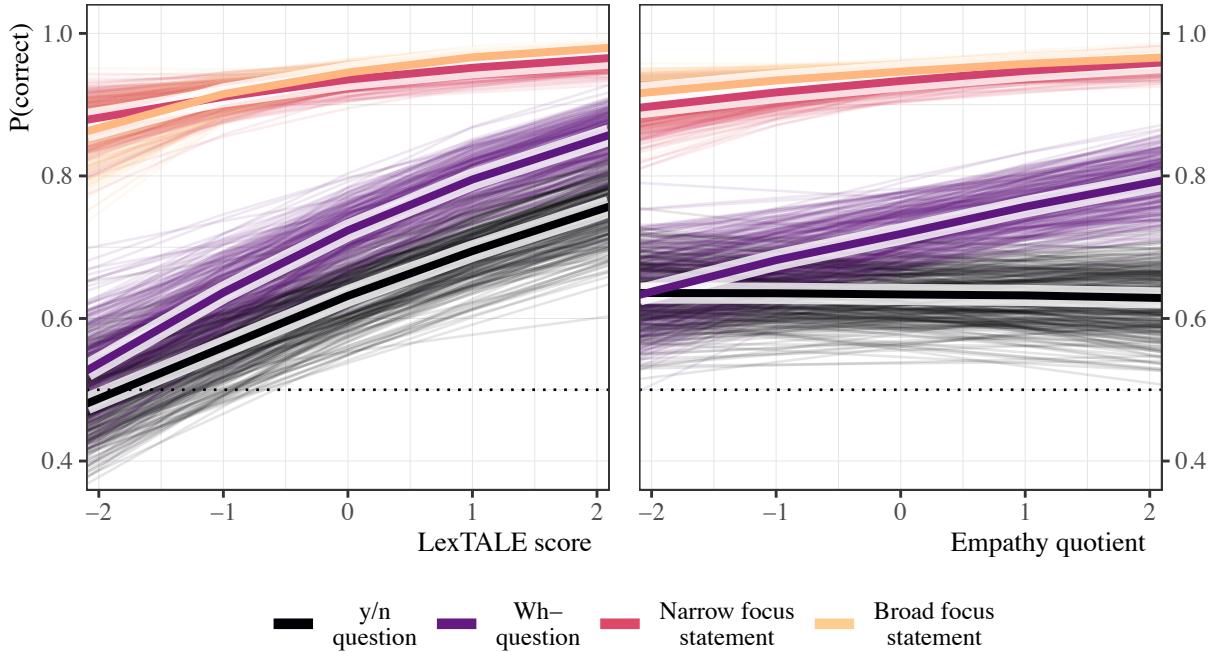


Figure 3. Conditional effects of a correct response as a function of proficiency (LexTALE score) (left panel) and empathy quotient (right panel) for each utterance type. Thin lines represent 300 draws from the posterior distribution for each condition and illustrate uncertainty (95% HDI) around the posterior medians (thick lines).

The omnibus model also estimated the proficiency \times empathy quotient simple interaction for each utterance type. We used the posterior distribution to estimate the probability that this effect was non-zero for each utterance type. We found evidence that the proficiency effect was modulated by empathy quotient scores for wh- questions ($\beta = 0.22$, HDI = $[0.05, 0.39]$, ROPE =

0.06 , MPE = 0.99), though not for yes/no questions ($\beta = 0.02$, HDI = $[-0.09, 0.14]$, ROPE = 0.93, MPE = 0.65), broad focus statements ($\beta = 0.10$, HDI = $[-0.14, 0.35]$, ROPE = 0.46, MPE = 0.80), nor narrow focus statements ($\beta = 0.04$, HDI = $[-0.17, 0.25]$, ROPE = 0.65, MPE = 0.64). This relationship is illustrated in Figure 4. Specifically, we plot conditional effects of response accuracy as a function of proficiency and empathy quotient for the yes/no and wh- questions. In the left panel of Figure 4, one observes a positive correlation between response accuracy and proficiency that remains constant at standardized empathy quotient values of -1 , 0 and $+1$ for the yes/no questions. For the wh- questions (right panel), on the other hand, we see that the slope of the proficiency effect increases for higher empathy quotient values. That is to say, for wh- questions, at a given proficiency level, learners with higher empathy (black lines) tended to respond more accurately.

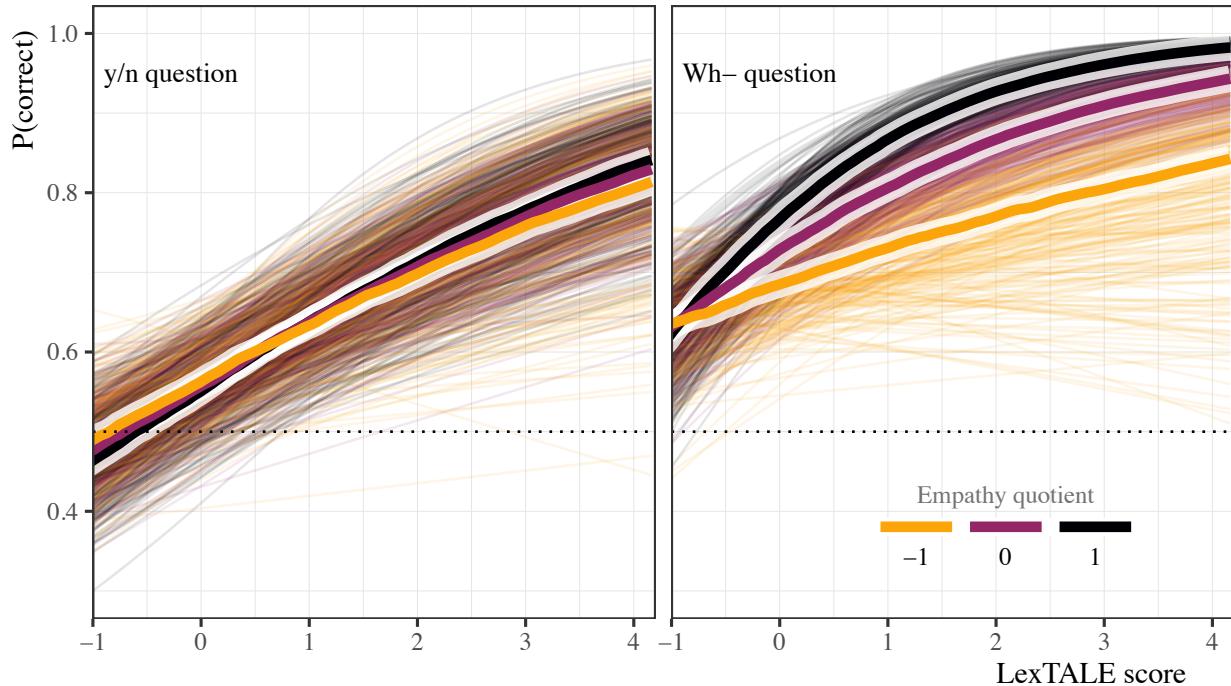


Figure 4. Probability of a correct response as a function of LexTALE score while holding empathy quotient scores constant at -1 , 0 and $+1$ standard deviations from the mean for each question type. Thin lines represent 300 draws from the posterior distribution and indicate uncertainty (95% HDI) around the posterior medians (thick lines).

With regard to response accuracy and response time differences based on speaker variety, we used the speaker variety grouping effect from the omnibus model to obtain posterior estimates (See Figure 5). As was the case with the monolingual Spanish pilot data, learners were least accurate when responding to the Cuban variety and most accurate when responding to the Peninsular-Madrileño and Mexican varieties. Response accuracy to a given variety did not correlate with response times. For instance, although learners were least accurate when responding to the Cuban stimuli, they had average response times similar to the grand mean for this variety.

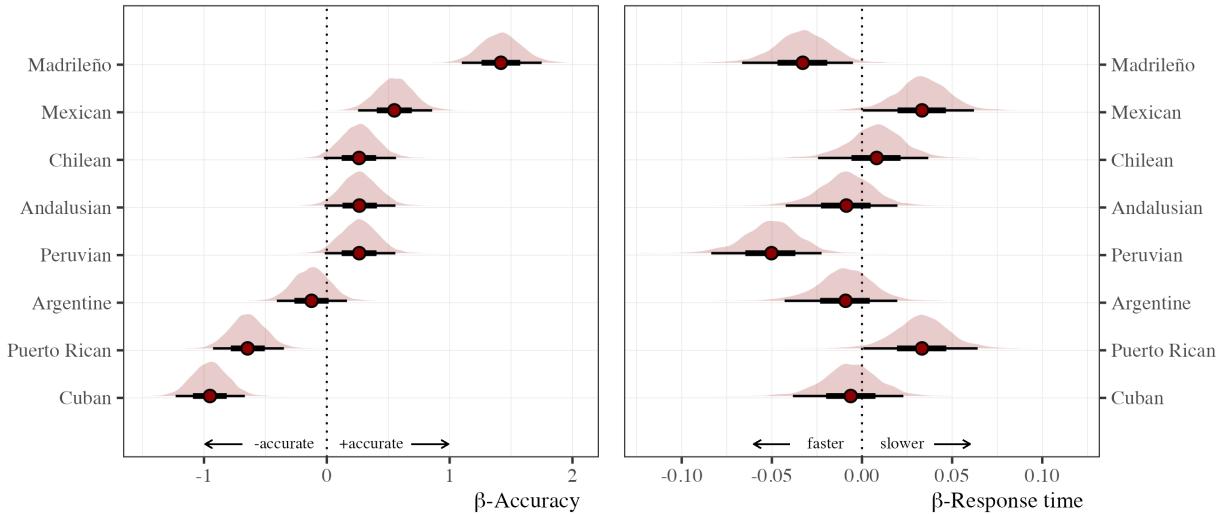


Figure 5. Grouping-level estimates of response accuracy and response time as a function of speaker variety. Red points represent posterior medians along with 66% and 95% highest density credible intervals. The vertical dotted lines indicate the grand mean.

Drift diffusion models

As described previously, we fit a drift diffusion model to each participants' data in order to obtain estimates for boundary separation (α) and drift rate (δ). Specifically, we fit two Bayesian measurement error models with the same functional form: boundary separation or drift rate as a function of utterance type, proficiency (LexTALE score), and empathy quotient. Given

the high accuracy on declarative statements, we focus our analyses on yes/no and wh- questions.

Figure 6 provides a forest plot summarizing the two models.

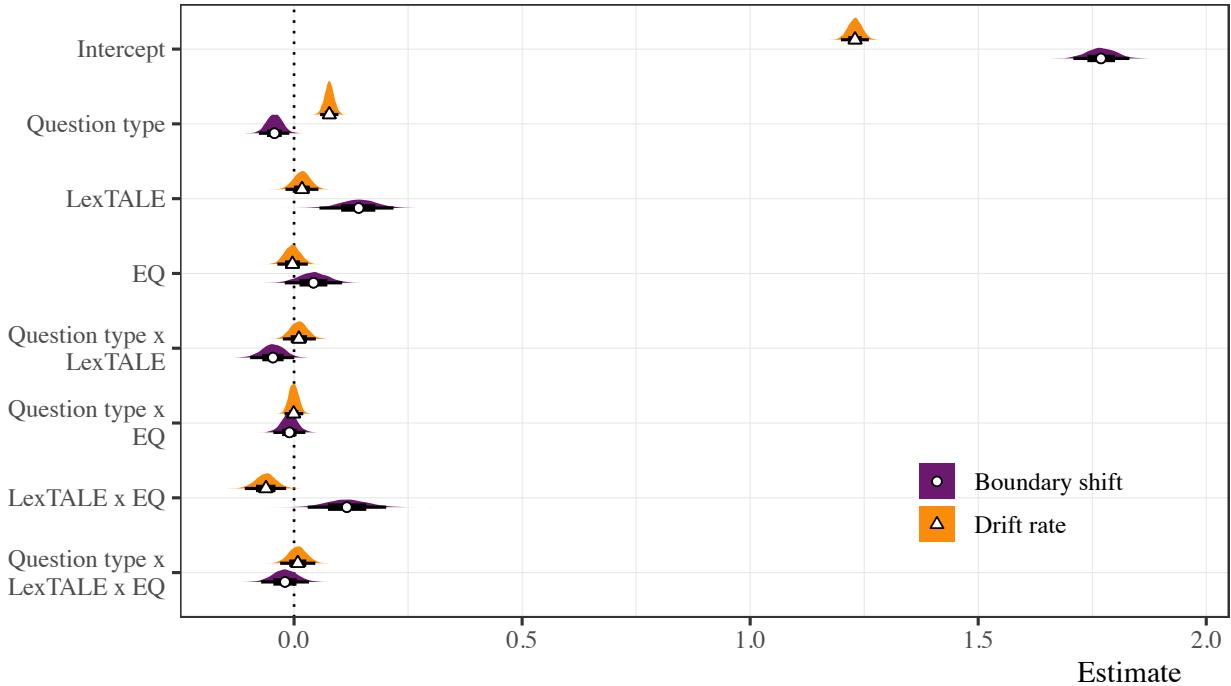


Figure 6. Forest plot summary of boundary separation (α , white circles under purple distributions) and drift rate (δ , white triangles under orange distributions) error measurement models.

Averaging over utterance type and holding proficiency and empathy quotient constant at the distribution means, posterior medians were positive for both boundary separation ($\beta = 1.77$, HDI = [1.70, 1.83], MPE = 1) and drift rate ($\beta = 1.23$, HDI = [1.20, 1.26], MPE = 1). Boundary separation was slightly lower in wh- questions ($\beta = -0.04$, HDI = [-0.08, -0.01], MPE = 0.99), suggesting that, overall, learners needed less information in order to make a decision when presented with questions of this type. Drift rate, on the other hand, was higher for wh- questions ($\beta = 0.08$, HDI = [0.06, 0.10], MPE = 1), which indicates that learners arrived at the decision threshold at a faster rate, and, thus, found this type of utterance to be easier. This corresponds with the finding that overall learners were more accurate responding to wh- questions than yes/no

questions by approximately 10% (Mean difference: $\beta = 9.30$, HDI = [3.74, 14.05], ROPE = 0.00, MPE = 1.00). Taken together, we can surmise that the ‘average’ learner has a lower threshold of required information in order to make a decision and arrives at this threshold at a faster rate for wh- questions in comparison with yes/no questions.

Crucially, in both models we also find evidence for a proficiency \times empathy quotient interaction. For both question types, boundary separation increased as a function of proficiency, but the association was conditional on empathy quotient score ($\beta = 0.12$, HDI = [0.03, 0.20], MPE = 1), with low empathy individuals seeing little to no change in estimated α . The effect was reversed for drift rate. In this case, estimated δ increased as a function of proficiency in low empathy individuals, and higher empathy individuals, particularly those with higher proficiency levels, saw decreases in drift rate ($\beta = -0.06$, HDI = [-0.11, -0.02], MPE = 1). To illustrate more clearly the practical relevance of these interactions, we ran 2,000 simulations from the drift diffusion model. Figure 7 plots the simulations for each question type at low/high proficiency and empathy levels (± 2 standard deviations). Individual lines represent random walks. The walk ends when enough evidence is accumulated and a decision threshold (horizontal, discontinuous grey lines) is reached. The upper threshold indicates a decision leading to a correct response and the lower threshold an incorrect response. Thick red lines indicate the simulation average for correct/incorrect responses in each condition. Focusing on the lower row of plots (high empathy), moving from left to right (low proficiency to high proficiency) within each question type, one observes (a) an increase in boundary separation (α), i.e. a greater distance between thresholds, via the horizontal grey lines, and (b) a decrease in drift rate (δ), i.e., a slower rate of information accumulation leading to a decision, via the horizontal distance of the red lines. In practical terms, this implies that high proficiency, high empathy learners required more information to reach a

decision and responded at a slower rate, compared to low empathy learners (top row), regardless of proficiency level.

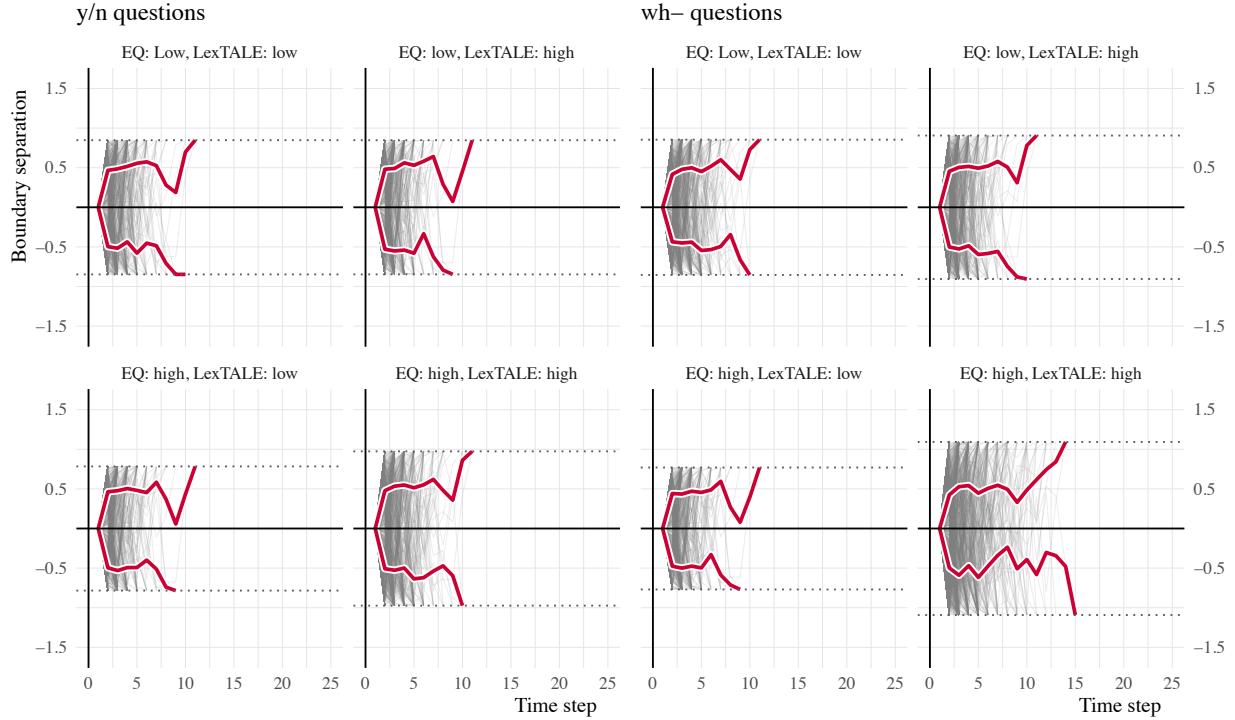


Figure 7. Two-thousand simulations of the drift diffusion model for interrogative utterances as a function of empathy quotient (low/high) and LexTALE score (low/high). Low and high levels represent ± 2 standard deviations above/below the mean. Horizontal, discontinuous grey lines indicate decision thresholds and dark red lines represent the simulation averages.

Discussion

The present work explored how the comprehension of intonation develops in adult L2 learners of Spanish. We used a two-alternative forced-choice task in which participants determined whether or not utterances presented in auditory stimuli were questions. Our study represents a conceptual replication of Brandl et al. (2020), but extends this research to address recent findings suggesting that individual pragmatic skill—in the context of the present work, empathy—plays a role in the process of forming sound-meaning associations. We used Bayesian methods, in particular Drift Diffusion modeling (Ratcliff & McKoon, 2008), to analyze data from

225 L2 learners. We find that perception and processing of intonation develops in tandem with proficiency in the target language and is, to some degree, modulated by the construct empathy.

This study set out to address three pre-registered research questions that we will now revisit.

The first question, *Is perceptual development in L2 Spanish modulated by proficiency and intonation type?*, was developed as a direct result of the previous literature examining the acquisition of Spanish prosody (i.e., Brandl et al., 2020; Trimble, 2013b). Response accuracy to all utterance types was positively correlated with proficiency, as measured by LexTALE scores. This corroborates the general finding that development of L2 intonation is positively correlated with target language proficiency, for both production (Craft, 2015; Henriksen et al., 2010; Thornberry, 2014; Trimble, 2013a, among others) and perception (Brandl et al., 2020; Nibert, 2005, 2006; Trimble, 2013b). In contrast with previous studies, our analyses conceptualized proficiency as a continuous variable, obviating the need to arbitrarily assign learners to proficiency groups. This operationalization will benefit future research interested in quantifying the effect of proficiency on perceptual development by allowing for more transparent designs with regard to statistical power and sample sizes. In line with previous studies (e.g., Brandl et al., 2020), we found that yes/no questions were most difficult for L2 learners of Spanish, followed by wh- questions and broad focus and narrow focus statements. An exploratory analysis using d' found that learner sensitivity to the utterance types followed the same pattern. While it is not clear exactly why yes/no questions are the most difficult, one possibility is that wh- questions pose less of a challenge because they contain a wh- word (e.g., *cuándo*, *cómo*, etc.). In other words, it might be the presence of a lexical cue in our task (and that of Brandl et al., 2020) that facilitates the interpretation of a wh- question in addition to intonation. At this juncture this possibility cannot be discarded, though it is worth noting that the presence of these words alone does not imply a question. That is to say, in specific contexts these same words can appear in statements as

well, in some cases with a pitch accent (i.e., *Qué beba María*) and in others without (i.e., *Que bebe María*). A particular intonation contour is typically present to force a question interpretation and said contour can vary between and even within varieties.⁴ Moreover, apart from the propositional content, a wh- question also implies a presupposition, and, thus, is more pragmatically complex. On the other hand, the yes/no questions in our experimental task have the same syntactic structure as the declarative statements. Perhaps for this reason yes/no questions require more effort and attention to intonation in order to distinguish them from statements in our task.

Additionally, our study addressed the question *Do pragmatic skills—specifically, empathy—modulate the rate of development in L2 prosody?* This question was motivated by a line of research showing that empathy influences language processing in monolingual populations (Esteve-Gibert et al., 2016, 2020; Orrico & D'Imperio, 2020). Though the construct *empathy* has been considered in the SLA literature, the current body of research is limited to studies on pronunciation accuracy (i.e., Guiora, Brannon, et al., 1972; Rota & Reiterer, 2009, among others). Thus, we extend research on empathy to L2 phonological acquisition as it relates to speech perception. Using a cross-sectional design, we show (1) that empathy, as measured by the empathy quotient (Baron-Cohen & Wheelwright, 2004), did indeed modulate response accuracy and the decision-making process, and (2) *how* empathy affected sentence processing was related to L2 proficiency. Specifically, we found response accuracy increased as a function of proficiency, independent of empathy for yes/no questions, but not wh- questions. In the case of the latter, we found empathy to have a compounding effect on the correlation between accuracy

⁴ See Supplementary Materials for more information regarding the intonation contours observed in the stimuli of the present work.

and proficiency, such that higher empathy individuals showed more accuracy at lower proficiency levels when compared with their lower empathy counterparts. This is taken as evidence suggesting that empathy can potentially modulate the rate of development of L2 prosody. In other words, higher empathy individuals may develop L2 prosody at an earlier stage than lower empathy individuals. That being said, we do not find the same effect with yes/no questions. This finding is quite puzzling, particularly because previous research on sentence processing has found an effect for empathy in yes/no questions, e.g., in Salerno Italian (Orrico & D'Imperio, 2020). At this time, we are uncertain as to why our results differ in this regard, though the nature of the outcome variable measured in the task used in Orrico and D'Imperio (2020) (certainty scores bounded at 0 and 100) may have provided a more fine-grained window into the effect of empathy.

In addition to addressing response accuracy, we also show that for high proficiency, high empathy learners (1) more information was necessary to reach a decision and (2) responses came at a slower rate when compared with low empathy learners at any proficiency level. This interaction effect on sentence processing was found for both types of interrogative utterances. Previous research on monolingual populations has shown that higher empathy individuals are more sensitive to intonation cues in the process of forming sound-meaning associations than lower empathy individuals. Our findings support the notion that this is also true for adult L2 learners, though we show that increased sensitivity does not necessarily imply increased processing speed. Given that empathy comprises the cognitive process of identifying the emotional state of another living being as well as the affective process of experiencing a similar sensation within oneself, it is plausible that higher empathy individuals showed more sensitivity to intonation cues and unconsciously devoted cognitive efforts to this process because they tended to require more information during decision-making. On the contrary, other individuals,

which did not require as much information for reaching a decision, likely did not employ the same cognitive and affective processes related to empathy.

Our third research question addressed the effect of speaker variety on L2 perceptual development. Specifically, we asked *Does speaker variety affect perception accuracy and processing speed?* This question was motivated by Brandl et al. (2020), who raised the possibility that dialectal or sociolectal variation could have influenced participants' responses in their data. Their study included stimuli from eight varieties of Spanish, though this factor was not considered in their analysis. Building on Brandl et al. (2020), our auditory stimuli also included eight distinct varieties of Spanish. We found that, generally, speaker variety did indeed affect response accuracy. As was the case with our pilot data from monolingual Spanish speakers, learners were most accurate responding to stimuli from the speaker of Peninsular-Madrileño Spanish, and least accurate when responding to the Cuban variety. Interestingly, accuracy with a given variety did not correlate with response times in a straightforward way. For instance, participants did not respond faster to the Peninsular-Madrileño variety even though they were more accurate in their responses to this speaker.

The results of our study suggest that speaker variety does affect perception accuracy, though this does not necessarily map directly on to processing speed. One possibility put forward in the literature is that the variety matters insomuch that it is familiar to the listener (see Perry, Mech, MacDonald, & Seidenberg, 2018; Trimble, 2013b). In other words, learners may be more accurate and process speech faster when listening to a variety they know well. Our study took into consideration familiarity, though the variety that was cited as being the most familiar, U.S.

Spanish (35% of 225 responses), was not one of the varieties presented in the stimuli.⁵ Mexican (21%) and Peninsular-Madrileño (20%) Spanish were reported as being the second and third most familiar varieties, and no participants indicated Cuban Spanish as being the variety they were most familiar with. To explore the effect of familiarity further, we conducted a non-preregistered analyses of the data from the participants who claimed to be most familiar with a Spanish variety that was included in our speaker varieties: Peninsular and Mexican Spanish.⁶ We coded the participants' responses to familiar versus unfamiliar varieties and fit a Bayesian logistic regression model to the data (See online supplementary materials for details). In short, we find that, marginalizing over proficiency and empathy, participants were indeed more accurate when responding to a familiar variety. This is true for all utterance types to a certain extent, but is more clearly the case for questions, likely because responses to declarative utterances were near ceiling. Figure 8 plots the familiarity effect for this subset of the data.

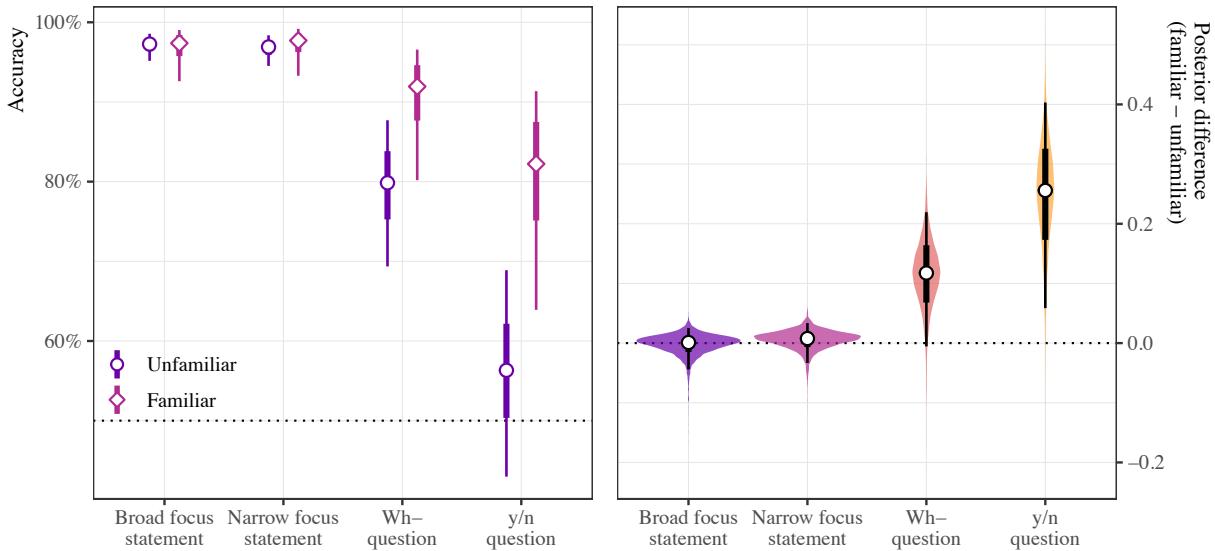


Figure 8. Response accuracy as a function of utterance type for unfamiliar and familiar Spanish

⁵ While participants mentioned familiarity with U.S. Spanish, it should be noted that this variety is not a monolith, but rather carries traits of the original Spanish variety (e.g., Mexican, Puerto Rican) that is in contact with English.

⁶ We make the assumption that 'Peninsular' is most closely associated with the Madrileño speaker.

varieties. Values represent posterior medians along with the 95% HDI for unfamiliar and familiar conditions (left panel), as well as the posterior difference (familiar - unfamiliar; right panel). The posterior predictive distribution is based on data from 91 participants who claimed to be familiar with Mexican ($n = 47$) and Peninsular ($n = 44$) Spanish.

Another plausible explanation for variety-specific difficulties lies in cross-linguistic differences in the prosodic realizations of the distinct utterance types. Yes/no questions in Peninsular-Madrileño Spanish, for example, have the common final rise found in many other varieties of Spanish, as well as Standard American English. Cuban and Puerto Rican Spanish, on the other hand, typically have a final fall (See Alvord, 2006; Armstrong, 2010, 2012; Hualde & Prieto, 2015; Sosa, 1999, among others). In our data, we do indeed find that L2 and native listeners are less accurate when responding to stimuli with final falls (See online supplementary materials for exploratory analyses), though these varieties were also considered to be less familiar. Ultimately our experimental design does not allow us to say definitively whether dialectal variation at the suprasegmental level accounts for variety-specific difficulties (as opposed to additional variation at the level of the segment, for example), though this reasoning is in line with previous studies, i.e., Trimble (2013b).

A final possibility is that speech rate differences associated with the speakers of the stimuli we used may have resulted in some varieties being more or less difficult for the learners (See Baese-Berk & Morrill, 2019). In an exploratory analysis of the auditory stimuli, we found that speech rate had no effect on response accuracy, as some of the varieties to which participants responded most accurately were also the fastest (e.g., the stimuli from our Mexican speaker). See Figure 13 of the supplementary materials for visualizations and further discussion.

In sum, the present work contributes to our knowledge of an understudied construct, empathy, as it pertains to speech. Additionally, this is the first time, to our knowledge, that drift

diffusion models have been used to analyze behavioral data relating to empathy in SLA. We also underscore the general need for models of L2 phonology, such as the SLM-r (Flege & Bohn, 2021), PAM-L2 (Best & Tyler, 2007), L2LP (Van Leussen & Escudero, 2015), etc., to address the acquisition process beyond the level of the segment. The LILT model (Mennen, 2015) has served as a starting point in the analysis of intonation across languages and L2 acquisition of intonation, framing the process of L2 acquisition of intonation along different developmental and structural dimensions, and has provided the theoretical grounding for numerous L2 studies (See Sánchez Alvarado & Armstrong, 2022; Sánchez-Alvarado, 2022, among others). The findings of the present study are in line with LILT since they show that perception of intonation in an L2 progresses with higher proficiency. In addition, these findings also emphasize the need for models like LILT to account for how individual differences pragmatic skills, such as empathy, can influence learner outcomes. A complete model of speech learning should account for both causal prediction and imputation at the segmental and suprasegmental levels. The present study aimed to address this gap in the literature by examining the role of proficiency and empathy on the perception of intonation during sentence processing in adult L2 phonological acquisition.

While the findings of our research suggest there is a relationship between target language proficiency and empathy, it is important to underscore that we do not make any claims about causality. Future research would benefit from considering the learnability of empathy (i.e., Bertrand, Guegan, Robieux, McCall, & Zenasni, 2018; Lam, Kolomitro, & Alamparambil, 2011) as it relates to L2 outcomes. Furthermore, the cross-sectional design of the present work is not ideal for addressing how empathy levels affect the rate at which perception of L2 intonation develops. Only longitudinal data can appropriately address this issue. On that note, at this time, research on speech perception and empathy is limited to intonation. A fruitful avenue for novel research ought to examine how empathy is related to perception and spoken word recognition at

the segmental level. A primary focus of the present project was to expand the line of research involving empathy and intonation perception in two ways: first, to individuals with different linguistic experience (specifically, L2 learners), and, second, to different communicative situations (utterance types). This project was not concerned with understanding *why* different pitch contours affect intonation perception, particularly with regard to the role of empathy, primarily because there is inherent variability in how speakers realize their communicative intentions, at both the variety and individual-level, within utterance types. This variability is also present in our stimuli. Future research would benefit from exploring *why* and *how* particular acoustic realizations of pitch within utterance types lead to distinct processing outcomes and how they might interact with pragmatic skill.

Conclusion

The present study investigated the development of L2 perception of intonation. Specifically, this study explored the relationship between target language proficiency and an individual pragmatic skill, empathy, in the process of learning Spanish as a second language by analyzing the perception of intonation in questions and statements. We find that perception of intonation in sentence processing develops in tandem with proficiency in the target language and interacts with individual empathy levels, supporting the general conclusion that higher empathic individuals, in comparison with lower empathic individuals, appear to be more sensitive to intonation cues in the process of forming sound-meaning associations. Importantly, increased sensitivity does not necessarily entail increased processing speed. The results motivate the inclusion of measures of pragmatic skill, such as empathy, to better account for intonational meaning processing and sentence comprehension in second language acquisition research.

References

- Alvord, S. M. (2006). *Spanish intonation in contact: The case of Miami Cuban bilinguals* (PhD thesis). University of Minnesota.
- Armstrong, M. E. (2010). Puerto Rican Spanish intonation. In P. Prieto & P. Roseano (Eds.), *Transcription of intonation of the Spanish language* (pp. 155–189). Münich: Lincom Europa.
- Armstrong, M. E. (2012). *The development of yes-no question intonation in Puerto Rican Spanish* (PhD thesis). The Ohio State University.
- Astruc, L., Vanrell, M., & Prieto, P. (2016). Cost of the action and social distance affect the selection of question intonation in Catalan. In M. E. Armstrong, N. Henriksen, & M. Vanrell (Eds.), *Intonational grammar in Ibero-Romance: Approaches across linguistic subfields* (pp. 93–113). John Benjamins Publishing Company.
<https://doi.org/10.1075/ihll.6>
- Aust, F., & Barth, M. (2018). *papaja: Create APA manuscripts with R Markdown*. Retrieved from <https://github.com/crsh/papaja>
- Aziz-Zadeh, L., Sheng, T., & Gheytanchi, A. (2010). Common premotor regions for the perception and production of prosody and correlations with empathy and prosodic ability. *PloS One*, 5(1), 1–8. <https://doi.org/10.1371/journal.pone.0008759>
- Baese-Berk, M. M., & Morrill, T. H. (2019). Perceptual consequences of variability in native and non-native speech. *Phonetica*, 76(2-3), 126–141. <https://doi.org/10.1159/000493981>

Baron-Cohen, S. (2011). *Zero degree of empathy. On empathy and the origins of cruelty.*

London, England: Penguin.

Baron-Cohen, S., & Wheelwright, S. (2004). The empathy quotient: An investigation of adults with Asperger syndrome or high functioning autism, and normal sex differences. *Journal of Autism and Developmental Disorders*, 34(2), 163–175.

<https://doi.org/10.1023/B:JADD.0000022607.19833.00>

Beckman, M. E., Díaz-Campos, M., McGory, J. T., & Morgan, T. A. (2002). Intonation across Spanish in the Tones and Break Indices framework. *Probus*, 9–36.

<https://doi.org/10.1515/prbs.2002.008>

Bertrand, P., Guegan, J., Robieux, L., McCall, C. A., & Zenasni, F. (2018). Learning empathy through virtual reality: Multiple strategies for training empathy-related abilities using body ownership illusions in embodied virtual reality. *Frontiers in Robotics and AI*, 5, 26.

<https://doi.org/10.3389/frobt.2018.00026>

Best, C. T., & Tyler, M. D. (2007). Nonnative and second-language speech perception: Commonalities and complementarities. In O.-S. Bohn & M. J. Munro (Eds.), *Language experience in second language speech learning: In honor of James Emil Flege* (pp. 13–34). Amsterdam, The Netherlands: John Benjamins.

Bishop, J. B. (2016). Individual differences in top-down and bottom-up prominence perception. *Proceedings of Speech Prosody*, 2016, 668–672.

- Bishop, J. B., Chong, A. J., & Jun, S.-A. (2015). Individual differences in prosodic strategies to sentence parsing. *Proceedings of the 18th International Congress of Phonetic Sciences*, 1–5. London: International Phonetic Association.
- Bishop, J. B., & Kuo, G. (2016). Do “autistic-like” personality traits predict prosody perception. *Talk Presented at LabPhon15 Satellite Workshop on Personality in Speech Perception and Production, Ithaca, NY*.
- Bishop, J. B., 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>
- Boersma, P., & Weenink, D. (2018). *Praat: Doing phonetics by computer [computer program]*. Retrieved from <http://www.praat.org/>
- Brandl, A., González, C., & Bustin, A. (2020). The development of intonation in L2 Spanish: A perceptual study. In A. Morales-Front, M. J. Ferreira, R. P. Leow, & C. Sanz (Eds.), *Hispanic linguistics: Current issues and new directions* (pp. 12–31). John Benjamins Publishing Company. <https://doi.org/10.1075/ihll.26>
- Brown, H. D. (1973). Affective variables in second language acquisition. *Language Learning*, 23(2), 231–244. <https://doi.org/10.1111/j.1467-1770.1973.tb00658.x>
- Bürkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, 80(1), 1–28. <https://doi.org/10.18637/jss.v080.i01>
- Bürkner, P.-C. (2018). Advanced Bayesian Multilevel Modeling with the R Package brms. *The R Journal*, 10(1), 395–411. <https://doi.org/10.32614/RJ-2018-017>

Butragueño, P. M. (2003). Hacia una descripción prosódica de los marcadores discursivos. Datos del español de México. *La Tonía: Dimensiones Fonéticas y Fonológicas*, 375–402.

Butragueño, P. M. (2004). Configuraciones circunflejas en la entonación del español mexicano. *Revista de Filología Española*, 84(2), 347–373.

Cain, K., Oakhill, J., & Lemmon, K. (2004). Individual differences in the inference of word meanings from context: The influence of reading comprehension, vocabulary knowledge, and memory capacity. *Journal of Educational Psychology*, 96(4), 671–681.

<https://doi.org/10.1037/0022-0663.96.4.671>

Carruthers, P. (2009). How we know our own minds: The relationship between mindreading and metacognition. *Behavioral and Brain Sciences*, 32(2), 121–138.

<https://doi.org/10.1017/S0140525X09000545>

Casielles-Suárez, E. (2004). *The syntax-information structure interface: Evidence from Spanish and English*. Routledge.

Chen, A. (2005). *On the universal and language-specific perception of paralinguistic intonational meaning* (PhD thesis). Utrecht: LOT.

Colantoni, L. (2011). *Broad-focus declaratives in Argentine Spanish contact and non-contact varieties*. John Benjamins.

Colantoni, L., & Gurlekian, J. (2004). Convergence and intonation: Historical evidence from Buenos Aires Spanish. *Bilingualism: Language and Cognition*, 7(2), 107–119.

<https://doi.org/10.1017/S1366728904001488>

Craft, J. (2015). *The acquisition of intonation by L2 Spanish speakers while on a six week study abroad program in Valencia, Spain* (PhD thesis). The Florida State University.

Cruz-Ferreira, M. (1987). Non-native interpretive strategies for intonational meaning: An experimental study. In A. James & J. Leather (Eds.), *Sound patterns in second language acquisition* (pp. 103–120). Berlin: Mouton de Gruyter.

Degen, J., & Tanenhaus, M. K. (2016). Availability of alternatives and the processing of scalar implicatures: A visual world eye-tracking study. *Cognitive Science*, 40(1), 172–201.

<https://doi.org/10.1111/cogs.12227>

de-la-Mota, C. (2019). Improving non-native pronunciation: Teaching prosody to learners of Spanish as a second/foreign language. In R. Rao (Ed.), *Key issues in the teaching of Spanish pronunciation* (pp. 162–197). Routledge.

Derwing, T. M., & Munro, M. J. (2015). *Pronunciation fundamentals: Evidence-based perspectives for L2 teaching and research* (Vol. 42). John Benjamins Publishing Company.

Diehl, J. J., Bennetto, L., Watson, D., Gunlogson, C., & McDonough, J. (2008). Resolving ambiguity: A psycholinguistic approach to understanding prosody processing in high-functioning autism. *Brain and Language*, 106(2), 144–152.

<https://doi.org/10.1016/j.bandl.2008.04.002>

Erteschik-Shir, N. (2007). *Information structure: The syntax-discourse interface* (Vol. 3). OUP Oxford.

- Estebas-Vilaplana, E., & Prieto, P. (2010). Castilian Spanish intonation. In P. Prieto & P. Roseano (Eds.), *Transcription of intonation of the Spanish language* (pp. 17–48). Münich: Lincom Europa.
- Esteve-Gibert, N., Portes, C., Schafer, A., Hemforth, B., & D'Imperio, M. (2016). The role of individual empathic skills on the online processing of intonational meaning. *Architectures and Mechanisms for Language Processing (AMLaP)*. Bilbao, Spain: Basque Center on Cognition, Brain; Language. <https://doi.org/10.13140/RG.2.2.19401.13926>
- Esteve-Gibert, N., Schafer, A. J., Hemforth, B., Portes, C., Pozniak, C., & D'Imperio, M. (2020). Empathy influences how listeners interpret intonation and meaning when words are ambiguous. *Memory & Cognition*, 48, 566–580. <https://doi.org/10.3758/s13421-019-00990-w>
- Face, T. L. (2003). Intonation in Spanish declaratives: Differences between lab speech and spontaneous speech. *Catalan Journal of Linguistics*, 2, 115–131. <https://doi.org/10.5565/rev/catjl.46>
- Face, T. L., & Prieto, P. (2007). Rising accents in Castilian Spanish: A revision of Sp_ToBI. *Journal of Portuguese Linguistics*, 6, 117–146. <https://doi.org/10.5334/jpl.147>
- Flege, J. E., & Bohn, O.-S. (2021). The revised speech learning model (SLM-r). In R. Wayland (Ed.), *Second language speech learning: Theoretical and empirical progress* (pp. 3–83). Cambridge University Press. <https://doi.org/10.1017/9781108886901.002>

- Frith, U., & Frith, C. D. (2003). Development and neurophysiology of mentalizing. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 358(1431), 459–473. <https://doi.org/10.1098/rstb.2002.1218>
- Gabriel, C., Feldhausen, I., Pešková, A., Colantoni, L., Lee, S., Arana, V., & Labastía, L. (2010). Argentinian Spanish intonation. In P. Prieto & P. Roseano (Eds.), *Transcription of intonation of the Spanish language* (pp. 285–317). Münich: Lincom Europa.
- Gelman, A., Simpson, D., & Betancourt, M. (2017). The prior can often only be understood in the context of the likelihood. *Entropy*, 19(10), 1–13. <https://doi.org/10.3390/e19100555>
- Guiora, A. Z., & Acton, W. R. (1979). Personality and language behavior: A restatement. *Language Learning*, 29(1), 193–204. <https://doi.org/10.1111/j.1467-1770.1979.tb01059.x>
- Guiora, A. Z., Beit-Hallahmi, B., Brannon, R. C., Dull, C. Y., & Scovel, T. (1972). The effects of experimentally induced changes in ego states on pronunciation ability in a second language: An exploratory study. *Comprehensive Psychiatry*, 13(5), 421–428. [https://doi.org/10.1016/0010-440X\(72\)90083-1](https://doi.org/10.1016/0010-440X(72)90083-1)
- Guiora, A. Z., Brannon, R. C., & Dull, C. Y. (1972). Empathy and second language learning. *Language Learning*, 22(1), 111–130. <https://doi.org/10.1111/j.1467-1770.1972.tb00077.x>
- Guiora, A. Z., Taylor, L., & Brandwin, M. (1968). The role of empathy in second language behavior. In *Proceedings of the 16th International Congress of Applied Psychology*. Amsterdam: Swets and Zeitlinger, 181–186.
- Henriksen, N., Armstrong, M. E., & García-Amaya, L. J. (2016). The intonational meaning of polar questions in Manchego Spanish spontaneous speech. In M. E. Armstrong, N.

Henriksen, & M. del M. Vanrell (Eds.), *Intonational grammar in Ibero-Romance: Approaches across linguistic subfields* (pp. 181–205). John Benjamins Publishing Company. <https://doi.org/10.1075/ihll.6>

Henriksen, N., & García-Amaya, L. J. (2012). Transcription of intonation of Jerezano Andalusian Spanish. *Estudios de Fonética Experimental*, 21, 109–162.

Henriksen, N., Geeslin, K. L., & Willis, E. W. (2010). The development of L2 Spanish intonation during a study abroad immersion program in León, Spain: Global contours and final boundary movements. *Studies in Hispanic and Lusophone Linguistics*, 3(1), 113–162.
<https://doi.org/10.1515/shll-2010-1067>

Hu, X., Ackermann, H., Martin, J. A., Erb, M., Winkler, S., & Reiterer, S. M. (2013). Language aptitude for pronunciation in advanced second language (L2) learners: Behavioural predictors and neural substrates. *Brain and Language*, 127(3), 366–376.
<https://doi.org/10.1016/j.bandl.2012.11.006>

Hualde, J. I., & Prieto, P. (2015). Intonational variation in Spanish: European and American varieties. In S. Frota & P. Prieto (Eds.), *Intonation in Romance* (pp. 350–391). Oxford University Press.

Izura, C., Cuetos, F., & Brysbaert, M. (2014). LexTALE-Esp: A test to rapidly and efficiently assess the Spanish vocabulary size. *Psicológica*, 35(1), 49–66.
<https://doi.org/10.1037/t47086-000>

Jilka, M. (2000). Testing the contribution of prosody to the perception of foreign accent. In A. James & J. Leather (Eds.), *In proceedings of new sounds 4th international symposium on*

the acquisition of second language speech. Amsterdam: University of amsterdam (Vol. 4, pp. 199–207).

Kruschke, J. K. (2018). Rejecting or accepting parameter values in Bayesian estimation.

Advances in Methods and Practices in Psychological Science, 1(2), 270–280.

Kvavik, K. H., & Olsen, C. L. (1974). Theories and methods in spanish intonational studies.

Phonetica, 30(2), 65–100. <https://doi.org/10.1159/000259481>

Labastía, L. O. (2006). Prosodic prominence in Argentinian Spanish. *Journal of Pragmatics*, 38(10), 1677–1705. <https://doi.org/10.1016/j.pragma.2005.03.019>

Labastía, L. O. (2011). Procedural encoding and tone choice in Buenos Aires Spanish.

Procedural Meaning: Problems and Perspectives, CRISPI, 25, 383–413.

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

Lam, T. C. M., Kolomitro, K., & Alamparambil, F. C. (2011). Empathy training: Methods, evaluation practices, and validity. *Journal of Multidisciplinary Evaluation, 7*(16), 162–200.

Lemhöfer, K., & Broersma, M. (2012). Introducing LexTALE: A quick and valid lexical test for advanced learners of English. *Behavior Research Methods, 44*(2), 325–343.

<https://doi.org/10.3758/s13428-011-0146-0>

Levis, J. (2016). Accent in second language pronunciation research and teaching. *Journal of Second Language Pronunciation, 2*(2), 153–159. <https://doi.org/10.1075/jslp.2.2.01lev>

Liu, Y. (2017). Study on the influence of emotion factors in Second Language Acquisition. In G. Yu, G. Ke, & L. Han (Eds.), *Proceedings of the International Conference on Financial Management, Education and Social Science* (pp. 261–264).

<https://doi.org/10.25236/fmess.2017.55>

Marasco, O. M. (2020). “*Are you asking me or telling me?*” *Perception and production of Y/N questions and statements in L2 Spanish* (PhD thesis). University of Toronto.

Mennen, I. (2007). Phonological and phonetic influences in non-native intonation. In J. Trouvain & U. Gut (Eds.), *Non-native prosody. Phonetic description and teaching practice* (pp. 53–76). Berlin: De Gruyter Mouton.

Mennen, I. (2015). Beyond segments: Towards a L2 intonation learning theory. In *Prosody and language in contact* (pp. 171–188). Springer. https://doi.org/10.1007/978-3-662-45168-7_9

Munro, M. J. (1995). Nonsegmental factors in foreign accent: Ratings of filtered speech. *Studies in Second Language Acquisition*, 17(1), 17–34.

<https://doi.org/10.1017/S0272263100013735>

Nibert, H. J. (2005). The acquisition of the phrase accent by intermediate and advanced adult learners of Spanish as a second language. *Selected Proceedings of the 6th Conference on the Acquisition of Spanish and Portuguese as First and Second Languages*, 108–122. Somerville, MA: Cascadilla Proceedings Project.

Nibert, H. J. (2006). The acquisition of the phrase accent by beginning adult learners of Spanish as a second language. *Selected Proceedings of the 2nd Conference on Laboratory*

Approaches to Spanish Phonetics and Phonology, 131–148. Somerville, MA: Cascadilla Proceedings Project.

Nieuwland, M. S., Ditman, T., & Kuperberg, G. R. (2010). On the incrementality of pragmatic processing: An ERP investigation of informativeness and pragmatic abilities. *Journal of Memory and Language*, 63(3), 324–346. <https://doi.org/10.1016/j.jml.2010.06.005>

O'Rourke, E. (2005). *Intonation and language contact: A case study of two varieties of Peruvian Spanish* (PhD thesis). University of Illinois at Urbana-Champaign.

Orrico, R., & D'Imperio, M. (2020). Individual empathy levels affect gradual intonation-meaning mapping: The case of biased questions in Salerno Italian. *Laboratory Phonology: Journal of the Association for Laboratory Phonology*, 11(1), 1–39.
<https://doi.org/10.5334/labphon.238>

Ortiz-Lira, H. (2003). Los acentos tonales en un corpus de Español de Santiago de Chile: Su distribución y realización. *La Tonía: Dimensiones Fonéticas y Fonológicas*, 303–316.

Ortiz-Lira, H., & Cid-Uribe, M. E. (2000). La prosodia de las preguntas indagativas y no-indagativas del Español culto de Santiago de Chile. *LEA: Lingüística Española Actual*, 22(1), 23–49.

Peirce, J. W., Gray, J. R., Simpson, S., MacAskill, M. R., Höchenberger, R., Sogo, H., ...

Lindelv, J. (2019). PsychoPy2: Experiments in behavior made easy. *Behavior Research Methods*, 51(1), 195–203.

Perry, L. K., Mech, E. N., MacDonald, M. C., & Seidenberg, M. S. (2018). Influences of speech familiarity on immediate perception and final comprehension. *Psychonomic Bulletin & Review*, 25(1), 431–439. <https://doi.org/10.3758/s13423-017-1297-5>

Pettorino, M., De Meo, A., & Vitale, M. (2014). Transplanting vowels towards the acoustic correlates of foreign accent. In Y. Congosto, M. L. Montero Curiel, & A. Salvador Plans (Eds.), *Fonética experimental, educación superior e investigación: II. Adquisición y aprendizaje de lenguas/español como lengua extranjera* (pp. 93–106).

Pickering, L. (2001). The role of tone choice in improving ITA communication in the classroom. *Tesol Quarterly*, 35(2), 233–255. <https://doi.org/10.2307/3587647>

Pierrehumbert, J. (1980). *The phonology and phonetics of English intonation* (PhD thesis). Massachusetts Institute of Technology.

Pierrehumbert, J., & Hirschberg, J. (1990). The meaning of intonational contours in the interpretation of discourse. In P. R. Cohen, J. L. Morgan, & M. E. Pollack (Eds.), *Intentions in communication* (pp. 271–312). MIT press.

Plonsky, L., & Oswald, F. L. (2014). How big is “big”? Interpreting effect sizes in L2 research. *Language Learning*, 64(4), 878–912. <https://doi.org/10.1111/lang.12079>

Quilis, A. (1981). *Fonética acústica de la lengua española*. Madrid: Gredos.

Quilis, A. (1987). Entonación dialectal hispánica. In H. López Morales & M. Vaquero (Eds.), *Actas del I Congreso internacional sobre el español de América. San Juan, Puerto rico: Academia puertorriqueña de la lengua española* (pp. 117–163).

Quilis, A. (1993). *Tratado de fonología y fonética españolas*. Madrid: Gredos.

Rao, R. (2019). Introduction. In R. Rao (Ed.), *Key issues in the teaching of Spanish pronunciation* (pp. 1–13). Routledge.

Ratcliff, R., & McKoon, G. (2008). The diffusion decision model: Theory and data for two-choice decision tasks. *Neural Computation*, 20(4), 873–922.

<https://doi.org/10.1162/neco.2008.12-06-420>

Rota, G., & Reiterer, S. M. (2009). Cognitive aspects of pronunciation talent. In G. Dogil & S. M. Reiterer (Eds.), *Language talent and brain activity* (pp. 67–96). Mouton de Gruyter.

<https://doi.org/10.1515/9783110215496>

Sánchez Alvarado, C., & Armstrong, M. (2022). Prosodic marking of object focus in L2 Spanish. *Studies in Hispanic and Lusophone Linguistics*, 15(1), 211–250.

<https://doi.org/10.1515/shll-2022-2060>

Sánchez-Alvarado, C. (2022). The acquisition of L2 Spanish intonation: An analysis based on features. *Journal of Second Language Pronunciation*, 8(1), 40–67.

<https://doi.org/10.1075/jslp.20041.san>

Shang, P. (2022). The role of native language experience and individual features in the cross-linguistic perception of Spanish intonation. *Proceedings of the 10th International Symposium on the Acquisition of Second Language Speech*.

<https://doi.org/10.13140/RG.2.2.13005.31201>

Sosa, J. M. (1999). *La entonación del español*. Madrid: Cátedra.

Sosa, J. M. (2003a). La notación tonal del español en el modelo Sp-ToBI. In P. Prieto (Ed.), *Teorías de la entonación* (pp. 185–208). Ariel.

- Sosa, J. M. (2003b). Wh-questions in Spanish: Meanings and configuration variability. *Catalan Journal of Linguistics*, 2, 229–247. <https://doi.org/10.5565/rev/catjl.51>
- Thornberry, P. A. (2014). *The L2 acquisition of Buenos Aires Spanish intonation during a study abroad semester* (PhD thesis). University of Minnesota.
- Trimble, J. C. (2013a). *Acquiring variable L2 Spanish intonation in a study abroad context* (PhD thesis). University of Minnesota.
- Trimble, J. C. (2013b). Perceiving intonational cues in a foreign language: Perception of sentence type in two dialects of Spanish. In C. Howe (Ed.), *Selected Proceedings of the 15th Hispanic Linguistics Symposium* (pp. 78–92). Somerville, MA: Cascadilla Proceedings Project.
- Trofimovich, P., & Baker, W. (2006). Learning second language suprasegmentals: Effect of L2 experience on prosody and fluency characteristics of L2 speech. *Studies in Second Language Acquisition*, 28(1), 1–30. <https://doi.org/10.1017/S0272263106060013>
- Van Leussen, J.-W., & Escudero, P. (2015). Learning to perceive and recognize a second language: The L2LP model revised. *Frontiers in Psychology*, 6, 6–12. <https://doi.org/10.3389/fpsyg.2015.01000>
- VanPatten, B. (2020). Input processing in adult L2 acquisition. In B. VanPatten, G. D. Keating, & S. Wulff (Eds.), *Theories in second language acquisition* (pp. 105–127). Routledge.
- Vilaplana Estebas, E., & Prieto, P. (2008). La notación prosódica del español: Una revisión del Sp_ToBI. *Estudios de Fonética Experimental*, 17, 264–283.

Wiener, S., & Bradley, E. D. (2020). Harnessing the musician advantage: Short-term musical training affects non-native cue weighting of linguistic pitch. *Language Teaching Research*. <https://doi.org/10.1177/13621688209717>

Willis, E. W. (2010). Dominican Spanish intonation. In P. Prieto & P. Roseano (Eds.), *Transcription of intonation of the Spanish language* (pp. 123–153). Münich: Lincom Europa.

Xu, Y. (2010). In defense of lab speech. *Journal of Phonetics*, 38(3), 329–336.

<https://doi.org/10.1016/j.wocn.2010.04.003>

Supplementary materials

Traditional analyses

This section contains additional information regarding the response accuracy and response time analyses, as well as tables reported but not included in the main text.

Learner response accuracy.

The population effects of the response accuracy model were specified in the following manner:

$$\begin{aligned} \text{is_correct}_{ij} &\sim \text{Bernoulli}(p_{ij}, m_{ij}) \\ \text{logit}(p_{ij}) &= \beta_0 + \beta_1 * \text{question_type} + \beta_2 * \text{LexTALE} + \beta_3 * \text{EQ} + \\ &\quad \beta_4 * \text{question_type} * \text{LexTALE} * \text{EQ} \end{aligned}$$

We employed the `0 + Intercept` syntax of `brms` and set weakly informative priors as follows:

$$\begin{aligned} \beta &\sim \text{Normal}(0, 0.3) \\ \sigma &\sim \text{Cauchy}(0, 0.1) \\ \rho &\sim \text{LKJcorr}(2) \end{aligned}$$

The summary of the response accuracy model is available in Table 2. The information provided in this table is equivalent to the left panel of Figure 2 in the manuscript.

Table 2: Summary of the posterior distribution modeling response accuracy as a function of utterance type, LexTALE, and Empathy quotient. The table includes posterior medians, the 95% HDI, the percentage of the HDI within the ROPE, and the maximum probability of effect (MPE).

Parameter	Median	HDI	% in ROPE	MPE	Rhat	ESS
Intercept	0.53	[0.23, 0.82]	0.00	1.00	1.00	2099
Wh- question	0.43	[0.17, 0.65]	0.00	1.00	1.00	2723
Narrow focus	2.13	[1.84, 2.37]	0.00	1.00	1.00	2608
Broad focus	2.34	[2.05, 2.59]	0.00	1.00	1.00	2576
LexTALE	0.28	[0.15, 0.41]	0.00	1.00	1.00	5245
Empathy quotient	-0.02	[-0.11, 0.09]	0.98	0.62	1.00	5351
Wh- question:LexTALE	0.12	[-0.06, 0.29]	0.42	0.90	1.00	4977
Narrow focus:LexTALE	0.02	[-0.17, 0.22]	0.72	0.58	1.00	8450
Broad focus:LexTALE	0.19	[-0.02, 0.42]	0.18	0.96	1.00	9482
Wh- question:EQ	0.20	[0.03, 0.36]	0.10	0.99	1.00	4912
Narrow focus:EQ	0.26	[0.08, 0.43]	0.02	1.00	1.00	8389
Broad focus:EQ	0.24	[0.05, 0.43]	0.05	0.99	1.00	8180
LexTALE:EQ	0.02	[-0.09, 0.14]	0.93	0.65	1.00	5954
Wh- question:LexTALE:EQ	0.19	[0.00, 0.39]	0.16	0.97	1.00	5604
Narrow focus:LexTALE:EQ	0.02	[-0.19, 0.23]	0.67	0.56	1.00	8344
Broad focus:LexTALE:EQ	0.08	[-0.17, 0.32]	0.51	0.74	1.00	8698

Drift diffusion models

Drift Diffusion Models (DDM), also referred to as Wiener Diffusion Models and Decision Diffusion Models, represent our preferred method for analyzing the data from our 2AFC task. DDMs are rarely used in SLA research, though they are commonplace in psychology. The primary selling point of using a DDM is related to the parameters the model estimates: boundary separation (α), drift rate (δ), bias (β), and non-decision time (τ). Together these parameters give rich information about the processes believed to underpin decision-making. Specifically, a DDM requires decision data, e.g., “left” or “right” choices, correct or incorrect responses, etc., and response times associated with said decisions. In linguistics, particularly in psycholinguistics, data of this nature derived from 2AFC tasks are often analyzed using separate models, one for responses, and another for response times (as we have done in our so-called ‘traditional analyses’).⁷ As mentioned, a DDM uses both of these dependent variables—responses and response times—to estimate the 4 aforementioned parameters. The estimates can then be scrutinized in subsequent models, if one estimates the parameters for each participant (i.e., the approach taken in the present work), and/or used for simulations. For our purposes, we employ the Bayesian implementation of the DDM, thus we sample from a posterior distribution of plausible estimates for α , δ , β , and τ for each participant. We then summarize and report these posterior distributions for statistical inferences.

The no-pooling models were fit using the following specification in `brms`:

```
rt_raw | dec(is_correct) ~ 0 + sentence_type,
bs ~ 0 + sentence_type,
ndt ~ 0 + sentence_type,
bias ~ 0 + sentence_type
```

and the priors were:

```
prior("normal(0, 1)", class = "b"),
prior("normal(0, 5)", class = "b", dpar = "bs"),
prior("normal(0.2, 1)", class = "b", dpar = "ndt"),
prior("normal(0.5, 1)", class = "b", dpar = "bias")
```

The complete code used to fit the models are available in `09_ddm.R` in the r scripts directory.

Measurement-error models. The measurement error models fit to the boundary separation and drift rate data were specified to include the standard error around each posterior median for α and δ :

$$\begin{aligned}\alpha &\sim Normal(\alpha_{n,TRUE}, SE_\alpha) \\ \delta &\sim Normal(\delta_{n,TRUE}, SE_\delta)\end{aligned}$$

⁷ Given how relatively uncommon DDMs are in linguistics, the present work includes both approaches, though it is reasonable to assume that this practice will diminish as DDMs become more well-known and the resources for implementing them become more user-friendly.

The priors for the drift rate model were:

$$\begin{aligned}\alpha &\sim \text{Normal}(1,0.5) \\ \beta &\sim \text{Normal}(0,0.3) \\ \tau &\sim \text{Cauchy}(0,0.3) \\ \sigma &\sim \text{Cauchy}(0,0.1) \\ \rho &\sim \text{LKJcorr}(2)\end{aligned}$$

and the priors for the boundary separation model were:

$$\begin{aligned}\alpha &\sim \text{Normal}(2,0.5) \\ \beta &\sim \text{Normal}(0,0.5) \\ \tau &\sim \text{Cauchy}(0,0.3) \\ \sigma &\sim \text{Cauchy}(0,0.1) \\ \rho &\sim \text{LKJcorr}(2)\end{aligned}$$

To specify this type of model in brms we use the `resp_se` function, as follows:

```
estimate | resp_se(se, sigma = TRUE) ~ 1 + # Criterion
          q_sum * lextale_std * eq_std +           # Population-Level effects
          (1 + q_sum * lextale_std * eq_std | participant) # Group-Level effects
```

The model summary is available in Table 3, which is equivalent to Figure 6 in the main document.

Table 3: Summary of the posterior distribution modeling boundary separation and drift rate as a function of question type, LexTALE, and Empathy quotient. The table includes posterior medians, the 95% HDI, the percentage of the HDI within the ROPE, and the maximum probability of effect (MPE).

Model	Parameter	Median	HDI	MPE	Rhat	ESS
Boundary separation	Intercept	1.77	[1.70, 1.83]	1.00	1.00	3407
	Question type	-0.04	[-0.08, -0.01]	0.99	1.00	3676
	LexTALE	0.14	[0.06, 0.22]	1.00	1.00	3460
	EQ	0.04	[-0.02, 0.11]	0.91	1.00	3585
	Question type:LexTALE	-0.05	[-0.09, 0.00]	0.97	1.00	3594
	Question type:EQ	-0.01	[-0.04, 0.03]	0.73	1.00	3993
	LexTALE:EQ	0.12	[0.03, 0.20]	1.00	1.00	3912
	Question type:LexTALE:EQ	-0.02	[-0.07, 0.03]	0.77	1.00	3580
	Drift rate Intercept	1.23	[1.20, 1.26]	1.00	1.00	3814
Drift rate	Question type	0.08	[0.06, 0.10]	1.00	1.00	3584
	LexTALE	0.02	[-0.02, 0.05]	0.83	1.00	3276
	EQ	0.00	[-0.03, 0.03]	0.59	1.00	4063
	Question type:LexTALE	0.01	[-0.02, 0.05]	0.70	1.00	3846
	Question type:EQ	0.00	[-0.02, 0.02]	0.53	1.00	4123
	LexTALE:EQ	-0.06	[-0.11, -0.02]	1.00	1.00	4114
	Question type:LexTALE:EQ	0.01	[-0.03, 0.05]	0.66	1.00	3733

Supplementary analyses

In this section we present supplementary analyses, all of which are exploratory in nature.

D'. Figure 9 and Table 4 represent an exploratory analysis of d' scores as a function of utterance type and speaker variety. One observes similar patterns to those from the accuracy analysis presented in the manuscript. The primary takeaway is that the analysis of learners' sensitivity to Spanish prosody mirrors that of their accuracy. That is to say, learners are more sensitive to (and accurate with) statements (broad focus, narrow focus) than questions (wh-, yes/no) (left panel of Figure 9). Learner sensitivity to the different Spanish varieties represented in the stimuli pattern in the same manner, i.e., more sensitivity to the Peninsular variety and less sensitivity to the Cuban variety (right panel of Figure 9). Table 4 summarizes the posterior of these exploratory analyses.

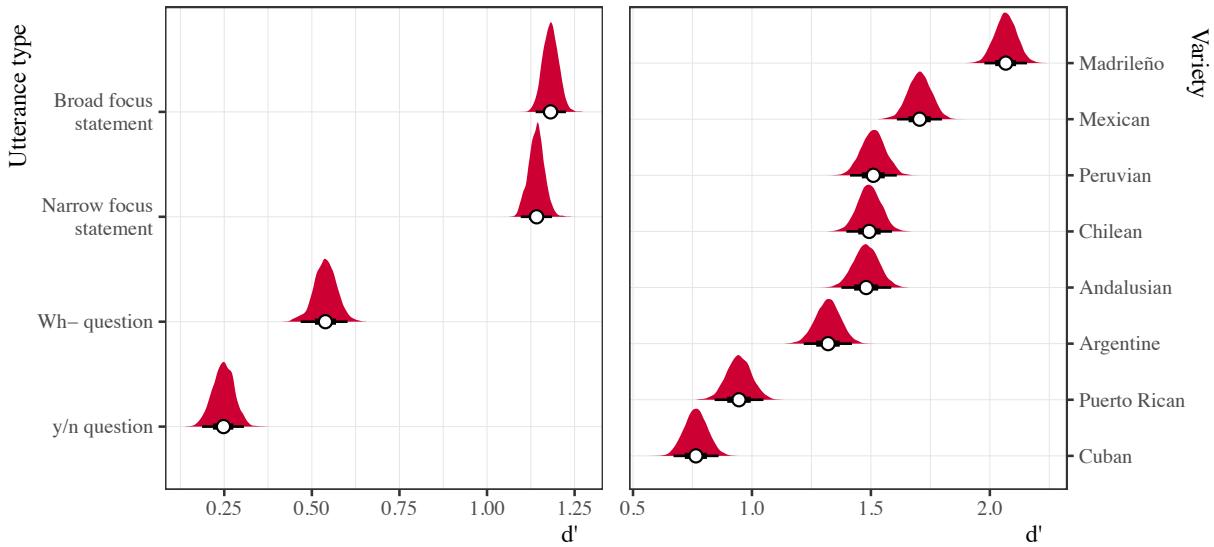


Figure 9. Exploratory analysis of d' as a function of utterance type and speaker variety. Points represent posterior medians with 66% and 95% credible intervals.

Table 4: Summary of the posterior distribution modeling d' as a function of question type or speaker variety. The table includes posterior medians, the 95% HDI, and the maximum probability of effect (MPE).

Model	Parameter	Median	HDI	MPE
Utterance type	Broad focus statement	1.18	[1.14, 1.23]	1.00
	Narrow focus statement	1.14	[1.10, 1.19]	1.00
	Wh- question	0.54	[0.47, 0.61]	1.00
	y/n question	0.25	[0.19, 0.31]	1.00
Variety	Andalusian	1.48	[1.38, 1.59]	1.00
	Argentine	1.32	[1.22, 1.42]	1.00
	Chilean	1.49	[1.40, 1.59]	1.00
	Cuban	0.76	[0.67, 0.86]	1.00
	Mexican	1.70	[1.61, 1.80]	1.00
	Peninsular	2.07	[1.98, 2.15]	1.00
	Peruvian	1.51	[1.42, 1.61]	1.00
	Puerto Rican	0.95	[0.85, 1.05]	1.00

Familiar vs. unfamiliar varieties. At the beginning of the experimental session, participants responded to the prompt ‘I am most familiar with Spanish from’ and, using a drop-down window, made a selection from the following choices: I am not familiar with Spanish, Argentina, Bolivia, Chile, Colombia, Costa Rica, Cuba, Dominican Republic, Ecuador, El Salvador, Equatorial Guinea, Guatemala, Honduras, Mexico, Nicaragua, Panama, Paraguay, Peru, Philippines, Puerto Rico, Uruguay, Venezuela, Spain, and United States. These choices represent the countries/territories/commonwealths where Spanish is spoken as an official or co-official language with the exception of Andorra, Belize, Gibraltar, and the Philippines, which were not included by mistake. Table 5 below summarizes the participants’ responses to this question. Of note, participants overwhelmingly selected ‘U.S. Spanish’, followed by ‘Mexico’, ‘Peninsular’, and ‘Not familiar’.

Table 5: Summary of participant self-reported familiarity with Spanish.

Variety	n	Proportion
U.S. Spanish	78	0.35
Mexico	47	0.21
Peninsular	44	0.20
Not familiar	38	0.17
Colombia	4	0.02
Costa Rica	4	0.02
Puerto Rico	4	0.02
Dominican Republic	2	0.01
Honduras	2	0.01
Peru	2	0.01

We did not pre-register a hypothesis regarding participant familiarity with Spanish. The following analysis is exploratory in nature. We analyzed the data from the participants who claimed to be most familiar with a Spanish variety that was included in our speaker varieties: Peninsular and Mexican Spanish (note: we make the assumption that ‘Peninsular’ is most closely associated with the Madrileño speaker). We coded the participants’ responses to familiar versus unfamiliar varieties and fit a Bayesian logistic regression model to the data. The model was specified similar to previous models:

$$\begin{aligned} \text{is_correct}_{ij} &\sim \text{Bernoulli}(p_{ij}, m_{ij}) \\ \text{logit}(p_{ij}) &= \beta_0 + \beta_1 * \text{question_type} + \beta_2 * \text{Familiarity} \\ &\quad \beta_3 * \text{question_type} * \text{Familiarity} \end{aligned}$$

We specified grouping variables for participant, speaker variety, and individual items. The models estimated varying slopes for the sentence type effect for each participant and the familiarity effect for each item. Again, we used the `0 + Intercept` syntax of brms and set weakly informative priors as follows:

$$\begin{aligned} \beta &\sim \text{Normal}(0,1) \\ \sigma &\sim \text{Cauchy}(0,0.2) \\ \rho &\sim \text{LKJcorr}(1) \end{aligned}$$

In short, we find that, marginalizing over proficiency and empathy, participants were more accurate when responding to a familiar variety. This is true for all utterance types to some degree, but more clearly the case for questions (likely because responses to declarative utterances were near ceiling). Figure 8 included in the manuscript illustrates the familiarity effect. Table 6 summarizes the model output. For convenience, we also provide the conditional effects of response accuracy in Table 7.

Table 6: Summary of the posterior distribution modeling response accuracy as a function of utterance type and familiarity. The model only includes data from participants who claimed to be familiar with Mexican ($n = 47$) and Peninsular ($n = 44$) Spanish. The table includes posterior medians, the 95% HDI, the percentage of the HDI within the ROPE, and the maximum probability of effect (MPE).

Parameter	Median	HDI	% in ROPE	MPE	Rhat	ESS
Intercept	0.25	[-0.28, 0.79]	0.19	0.84	1	1610
Wh- question	1.13	[0.71, 1.55]	0.00	1.00	1	2066
Narrow focus	3.19	[2.71, 3.71]	0.00	1.00	1	2561
Broad focus	3.32	[2.87, 3.82]	0.00	1.00	1	3217
Familiar	1.28	[0.26, 2.17]	0.00	0.99	1	2404
Wh- question:Familiar	-0.24	[-0.76, 0.32]	0.20	0.80	1	5970
Narrow focus:Familiar	-0.98	[-1.66, -0.24]	0.00	0.99	1	6516
Broad focus:Familiar	-1.24	[-1.94, -0.51]	0.00	1.00	1	5998

Table 7: Conditional effects of response accuracy as a function of sentence type and familiarity with the Spanish variety. Values represent posterior medians along with the 95% HDI for unfamiliar and familiar conditions, along with the posterior difference (familiar - unfamiliar). The posterior predictive distribution is based on data from participants who claimed to be familiar with Mexican ($n = 47$) and Peninsular ($n = 44$) Spanish.

Sentence type	Unfamiliar	Familiar	Difference
Broad focus statement	0.97 [0.95, 0.99]	0.97 [0.94, 0.99]	0.00 [-0.04, 0.03]
Narrow focus statement	0.97 [0.95, 0.99]	0.98 [0.95, 1.00]	0.01 [-0.03, 0.04]
Wh- question	0.80 [0.70, 0.88]	0.92 [0.83, 0.97]	0.12 [0.00, 0.22]
y/n question	0.56 [0.43, 0.69]	0.82 [0.67, 0.93]	0.26 [0.07, 0.41]

Rising vs. falling contours in yes/no questions. As pointed out by one of the anonymous reviewers, in our stimuli, the yes/no questions have rising intonation for all varieties except Cuban and Puerto Rican (see *Auditory stimuli* below). We calculated response accuracy as a function of intonational contour (falling: Cuban, Puerto Rican; rising: other) for yes/no questions using L2 and native listener data. Table 8 provides descriptive statistics. One can see that both listener types were least accurate when responding to the falling contours, though this is particularly true for the Cuban variety. This finding is illustrated in the previous section with regard to familiarity and is seen again in the following section regarding native listener data.

Table 8: L2 and native listener response accuracy to yes/no questions as a function of variety (Caribbean, Non-Caribbean). The stimuli from the Caribbean varieties include falling intonational contours.

Group	Cuban	Puerto Rican	Other
Native listeners	0.55 (0.5)	0.8 (0.4)	0.93 (0.25)
L2 learners	0.13 (0.33)	0.28 (0.45)	0.73 (0.44)

Similar to what was found in our primary analyses regarding yes/no questions, in this subset of the data, empathy had no effect on response accuracy regardless of whether the stimuli was produced with falling (i.e., Caribbean: Cuban, Puerto Rican) or rising (other) intonation (Figure 10).

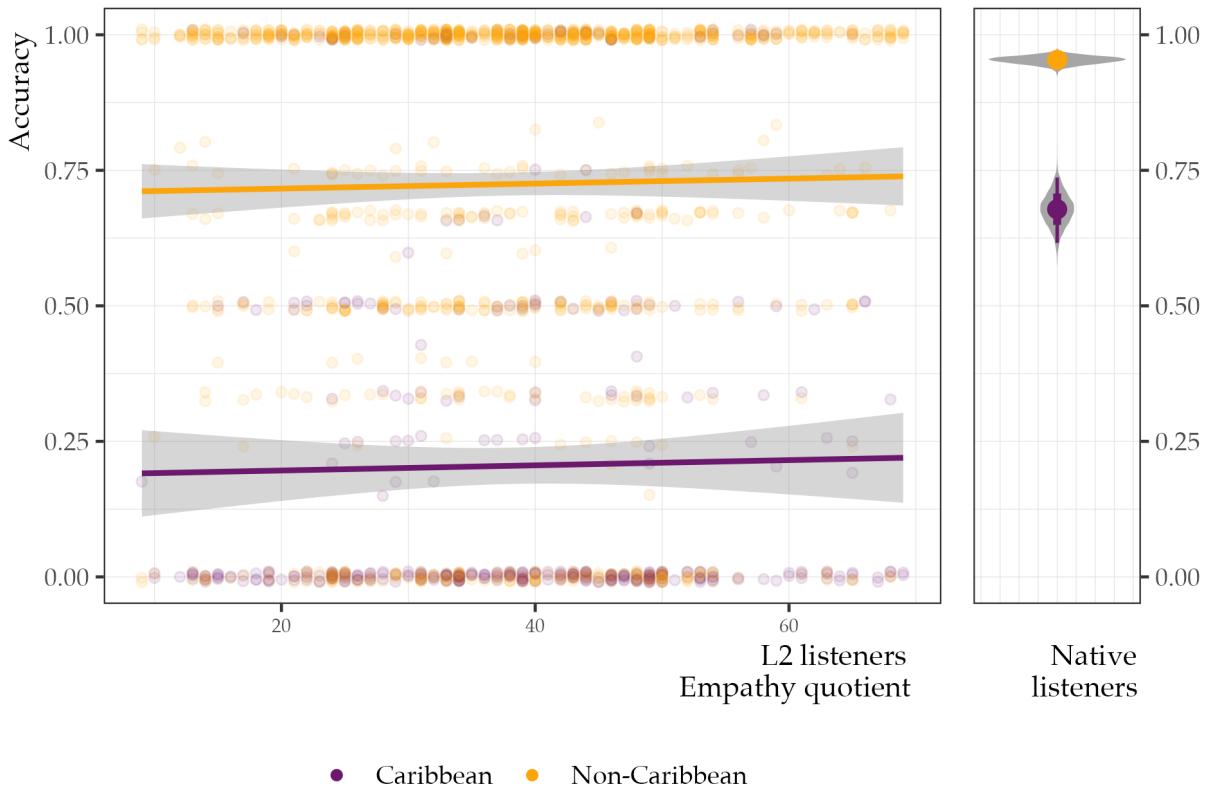


Figure 10. L2 and native listener response accuracy to Caribbean (Cuban, Puerto Rican) and non-Caribbean varieties for yes/no questions. Points represent by-participant means (learners) or posterior means (natives). For learners, the horizontal axis plots empathy quotient scores.

Monolingual response accuracy. In preparing our materials before official data collection, we piloted the 2AFC task and the auditory stimuli on monolingual Spanish speakers. The purpose of collecting this pilot data was to get an assessment of task difficulty—overall and as a function of speaker variety—and to have an idea what reasonable priors would be with regard to response times. From this pilot, we learned that, overall, monolingual speakers were least accuracy when

responding to the Cuban variety, and to some degree the Puerto Rican variety as well. This finding led us to hypothesize that L2 learners would also have difficulties when responding to stimuli from the same varieties. Figure 11 plots the monolingual accuracy data (right panel) next to the learner accuracy data (left panel). The same information is also provided in numeric form in Table 9.

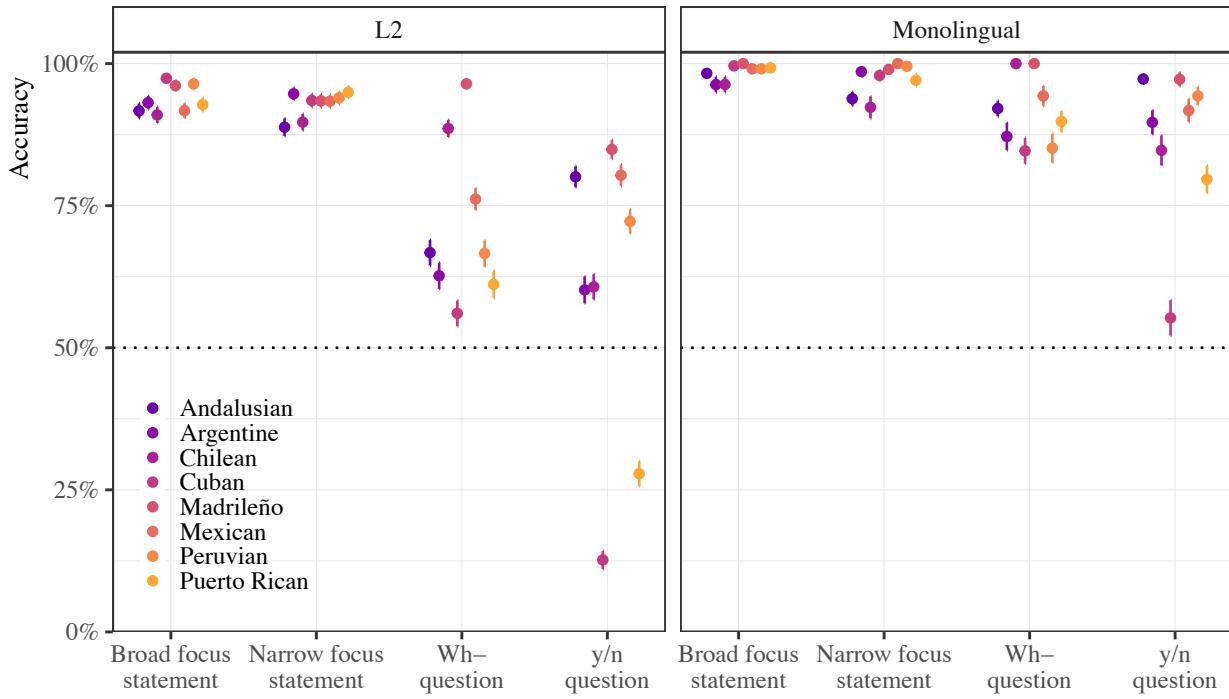


Figure 11. Response accuracy as a function of group (L2 learner, monolingual Spanish speaker), speaker variety (Andalusian, Argentine, Chilean, Cuban, Madrileño, Mexican, Peruvian, Puerto Rican), and utterance type (broad focus statement, narrow focus statement, wh- question, y/n question). Points represent means of the raw data surrounded by the standard error of the mean.

Table 9: Response accuracy as a function of group (L2 learner, monolingual Spanish speaker), speaker variety (Andalusian, Argentine, Chilean, Cuban, Madrileño, Mexican, Peruvian, Puerto Rican), and utterance type (broad focus statement, narrow focus statement, wh- question, y/n question). Each column provides the mean and standard error.

Type	Variety	L2	Monolingual
Broad focus statement	Andalusian	0.92 [0.90, 0.93]	0.98 [0.98, 0.99]
	Argentine	0.93 [0.92, 0.94]	0.96 [0.95, 0.98]
	Chilean	0.91 [0.90, 0.92]	0.96 [0.95, 0.98]
	Cuban	0.97 [0.97, 0.98]	1.00 [0.99, 1.00]
	Madrileño	0.96 [0.95, 0.97]	1.00 [1.00, 1.00]
	Mexican	0.92 [0.90, 0.93]	0.99 [0.98, 1.00]
	Peruvian	0.96 [0.96, 0.97]	0.99 [0.98, 1.00]
	Puerto Rican	0.93 [0.92, 0.94]	0.99 [0.99, 1.00]
Narrow focus statement	Andalusian	0.89 [0.87, 0.90]	0.94 [0.93, 0.95]
	Argentine	0.95 [0.94, 0.96]	0.99 [0.98, 0.99]
	Chilean	0.90 [0.88, 0.91]	0.92 [0.90, 0.94]
	Cuban	0.94 [0.92, 0.95]	0.98 [0.97, 0.99]
	Madrileño	0.93 [0.92, 0.95]	0.99 [0.98, 1.00]
	Mexican	0.93 [0.92, 0.95]	1.00 [1.00, 1.00]
	Peruvian	0.94 [0.93, 0.95]	1.00 [0.99, 1.00]
	Puerto Rican	0.95 [0.94, 0.96]	0.97 [0.96, 0.98]
Wh- question	Andalusian	0.67 [0.64, 0.69]	0.92 [0.91, 0.93]
	Argentine	0.63 [0.60, 0.65]	0.87 [0.85, 0.90]
	Chilean	0.89 [0.87, 0.90]	1.00 [1.00, 1.00]
	Cuban	0.56 [0.54, 0.58]	0.85 [0.82, 0.87]
	Madrileño	0.96 [0.96, 0.97]	1.00 [1.00, 1.00]
	Mexican	0.76 [0.74, 0.78]	0.94 [0.93, 0.96]
	Peruvian	0.67 [0.64, 0.69]	0.85 [0.83, 0.88]
	Puerto Rican	0.61 [0.59, 0.64]	0.90 [0.88, 0.92]
y/n question	Andalusian	0.80 [0.78, 0.82]	0.97 [0.96, 0.98]
	Argentine	0.60 [0.58, 0.63]	0.90 [0.88, 0.92]
	Chilean	0.61 [0.59, 0.63]	0.85 [0.82, 0.87]
	Cuban	0.13 [0.11, 0.14]	0.55 [0.52, 0.58]
	Madrileño	0.85 [0.83, 0.87]	0.97 [0.96, 0.98]
	Mexican	0.80 [0.78, 0.82]	0.92 [0.90, 0.94]
	Peruvian	0.72 [0.70, 0.74]	0.94 [0.93, 0.96]
	Puerto Rican	0.28 [0.26, 0.30]	0.80 [0.77, 0.82]

As a check on the low accuracy with the Cuban and Puerto Rican varieties, we decided to explore monolingual response accuracy further. To this end, we looked at the monolinguals' responses when they were presented with stimuli from their own variety, e.g., an Andalusian listener responding to stimuli from the Andalusian speaker. In our data, this implied a subset of Andalusian, Chilean, Cuban, Madrileño, Mexican, and Puerto Rican listeners. Figure 12 plots the variety-matched raw accuracy scores as a function of utterance type and Table 10 provides the same information in numeric form. Of note, all the monolinguals in our sample were at ceiling for all utterance types when responding to speakers from their own variety. This is taken as evidence that the auditory stimuli are accurate representations of questions and statements for these varieties.

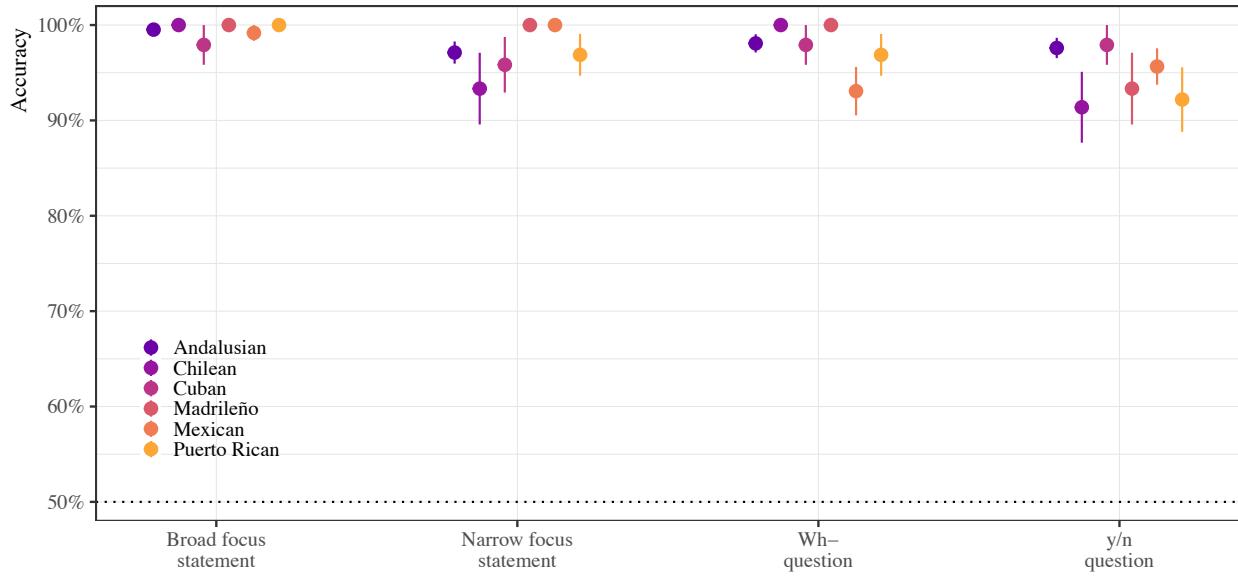


Figure 12. Variety-matched accuracy of monolingual listeners as a function of utterance type (broad focus statement, narrow focus statement, wh- question, y/n question). Points represent means of the raw data surrounded by the standard error of the mean.

Table 10: Variety-matched response accuracy as a function of utterance type. Accuracy refers to the proportion of correct responses along with the standard error of the mean.

Variety	Broad focus statement	Narrow focus statement	Wh- question	y/n question
Andalusian	1.00 [0.99, 1.00]	0.97 [0.96, 0.98]	0.98 [0.97, 0.99]	0.98 [0.97, 0.99]
Chilean	1.00 [1.00, 1.00]	0.93 [0.90, 0.97]	1.00 [1.00, 1.00]	0.91 [0.88, 0.95]
Cuban	0.98 [0.96, 1.00]	0.96 [0.93, 0.99]	0.98 [0.96, 1.00]	0.98 [0.96, 1.00]
Madrileño	1.00 [1.00, 1.00]	1.00 [1.00, 1.00]	1.00 [1.00, 1.00]	0.93 [0.90, 0.97]
Mexican	0.99 [0.98, 1.00]	1.00 [1.00, 1.00]	0.93 [0.91, 0.96]	0.96 [0.94, 0.98]
Puerto Rican	1.00 [1.00, 1.00]	0.97 [0.95, 0.99]	0.97 [0.95, 0.99]	0.92 [0.89, 0.96]

Auditory stimuli

Speakers. The auditory stimuli consisted of eight varieties of Spanish: Cuban, Peninsular-Madrileño, Peninsular-Andalusian, Puerto Rican, Chilean, Argentine, Mexican, and Peruvian. The speakers from Argentina, Chile, Peru, and Spain are linguists. Table 11 contains demographic information about the speakers.

Table 11: Demographic information for the eight varieties of Spanish represented in the auditory stimuli.

Country	City/Variety	Gender	Age
Argentina	Buenos Aires	Male	27
Chile	Valparaíso	Female	42
Cuba	Havana	Female	55
Mexico	Mexico City	Female	30
Peru	Lima	Male	30
Puerto Rico	Ponce	Female	35
Spain	Cádiz (Andalusia)	Female	35
Spain	Madrid	Female	29

Speech rate. In order to evaluate the possibility that the speech rate of the talkers in our stimuli may have affected response accuracy, we calculated the articulation rate (syllables spoken per second during phonation time), average syllable duration (in milliseconds), and speech rate (number of syllables divided by total time) for all items ($64 \text{ items} \times 8 \text{ speakers} = 512 \text{ utterances}$). These values are provided in Table 12.

Table 12: Average articulation rate (number of syllables divided by total phonation time), syllables duration (in milliseconds), and speech rate (number of syllable divided by total time) for each variety of the acoustic stimuli presented to listeners.

Variety	Articulation rate	Syllable duration	Speech rate
Andalusian	3.75	288	3.72
Argentine	3.59	300	3.59
Madrileño	3.72	282	3.72
Chilean	3.95	264	3.95
Cuban	3.99	272	3.99
Mexican	4.51	230	4.51
Peruvian	3.70	286	3.70
Puerto Rican	4.02	277	4.02

Figure 13 plots the posterior medians along with 66% and 95% HDI of standardized articulation rate for each variety. The plot shows that, generally, there was not a lot of variability between varieties. Though some of the slower varieties are also those in which we see higher response accuracy (compare with Figure 5), that was not always the case. That is, some of the faster varieties also had high response accuracy, e.g., Mexican Spanish.

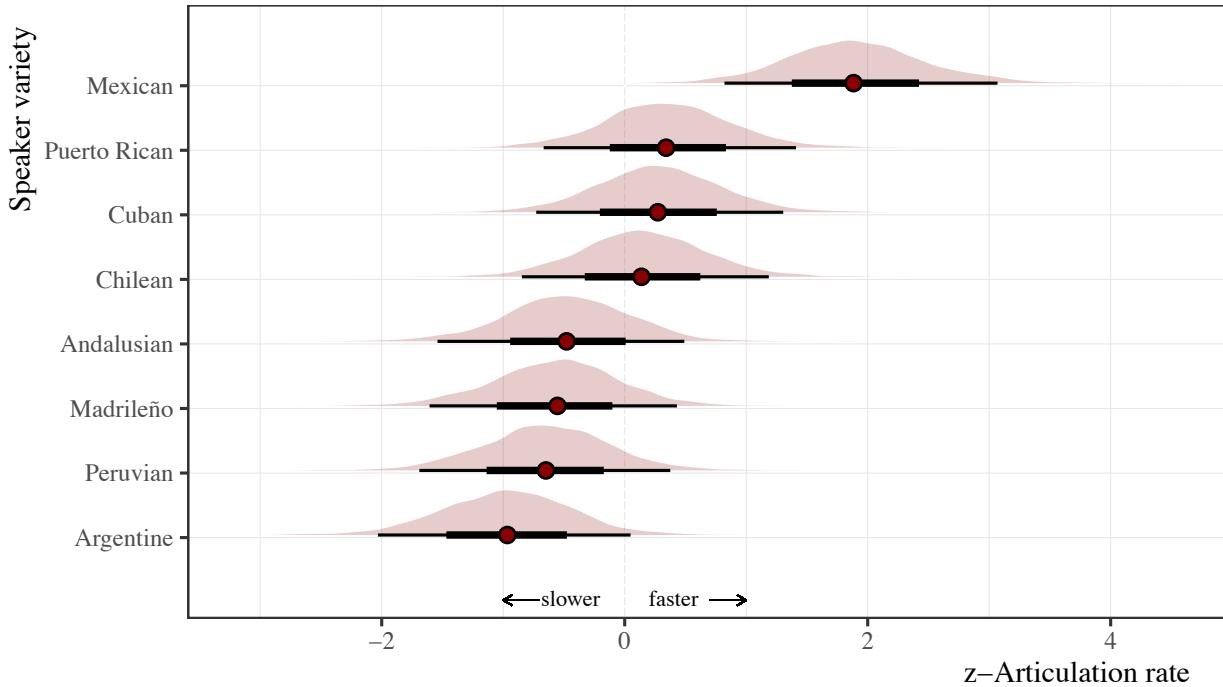


Figure 13. Standardized articulation rate as a function of speaker variety. Points represent posterior medians along with 66% and 95% HDI. The value ‘0’ on the horizontal axis represents the grand mean in the standardized space, which is equivalent to 3.90 syllables per second.

Acoustic description. In this subsection we provide examples of intonation contours for the four utterance types (broad focus state, narrow focus statement, wh- question, yes/no question) from the speakers of our eight varieties of Spanish (Cuban, Peninsular-Madrileño, Peninsular-Andalusian, Puerto Rican, Chilean, Argentine, Mexican, and Peruvian). Before proceeding to the acoustic descriptions, we will first provide an overview of the theoretical framework underpinning the descriptions.

The Autosegmental Metrical (AM) framework, developed by researchers like Pierrehumbert (1980) and Ladd (2008), aims to map the instrumentally-derived F0 values onto sequential, discrete and categorically-distinct intonational events. Within the AM framework, F0 is mapped onto pitch targets. Pitch targets can be monotonous, H(igh) and L(ow), or bitonal combinations of H and L, which represent rises (e.g., L+H*) or falls (e.g., H+L*). Pitch targets are identified at three points in an utterance: * indicates a pitch accent, which is associated and aligned with a stressed syllable; - indicates a phrase accent, which is associated with the boundary of an intermediate phrase; and % indicates a boundary tone or intonational phrase. It is important to note that an H at any prosodic level is the same H. The symbol that comes associated with it, whether *, -, or %, is merely giving information about how that tone is aligned with a specific event or boundary, not about the phonetic realization of that tone.

These abstract representations are mapped individually to instances of phonetic realizations specific to each language, similar to how /t/ has different acoustic properties in English and Spanish, as well as different phonetic realizations within a language depending on the context,

similar to allophones. This is further complicated by dialect, as different dialects may present different relationship mappings.

The Tones and Breaks Indices (ToBI) labeling system is used within the AM framework to annotate utterances for intonation. ToBI has two obligatory tiers: an orthographic tier, on which the utterance is recorded in IPA; and a tone tier, on which the pitch targets are recorded. A third, optional tier is the break-index tier, on which prosodic breaks and their relative strengths are recorded on a scale of 0 (no break) to 4 (Intonation Phrase break). It is idealized that each variety of a language should have its own ToBI system due to different varieties having distinct, but similar, intonational inventories. For example, in Dominican and Puerto Rican Spanish varieties, intonational phonologists may identify a bitonal fall H+L*, but the phonetic realization and pragmatic contexts in which that tone is produced may differ. For the present work we adopt the Spanish Tones and Breaks Indices (Sp_ToBI) as described in Hualde and Prieto (2015).

Using the recordings, we generated figures that display the waveform, spectrogram, and fundamental frequency contours. The figures are accompanied with two annotation tiers: an orthographic tier and a tone tier. We use this information together to identify associations between tonal targets and specific structure within example utterances following the Autosegmental Metrical model (Pierrehumbert, 1980). We follow the conventions of the Spanish Tones and Break Indices (Sp_ToBI) and its revisions (Beckman, Diaz-Campos, McGory, & Morgan, 2002; Face & Prieto, 2007; Sosa, 2003a; Vilaplana Estebas & Prieto, 2008). Our description is not meant to be exhaustive, but rather to give the interested reader an idea of how some of the contours included in our stimuli looked. Here we provide a single example of each utterance type from each speaker (Figures 14-21), though we encourage the interested reader to take a look/listen to all items, which are freely available on the OSF in the `data/stimuli` directory.

The statement utterances from our speaker of Madrid Spanish included a nuclear configuration of L* L%. The examples provided in the top row of Figure 14 are primarily differentiated by a prenuclear L>H* in the broad focus condition and L+H* in the narrow focus condition. Wh- and yes/no question had a similar prenuclear distinction, respectively. Our speaker also tended to use a final rise (L* ;H%) in wh- questions, which differs from the more common final fall. This is often associated with politeness (Quilis, 1993; Sosa, 2003b), or possibly “[...] a nuance of interest and greater speaker involvement in the speech act” (Estebas-Vilaplana & Prieto, 2010, p. 35). We believe we may have obtained more formal utterances, or possibly ‘lab’ speech due to the context in which the stimuli were recorded, i.e., in a lab.⁸ For more examples, see <http://prosodia.upf.edu/atlasentonacion/enquestes/espanol/madrid/index-english.html#maptask> (“¿Dónde está tu gasolina?”, minute: -3:18). For more description of Madrileño Spanish see also Quilis (1981), Quilis (1987), and Sosa (1999).

⁸ For more information on lab speech, see Face (2003) and Xu (2010).

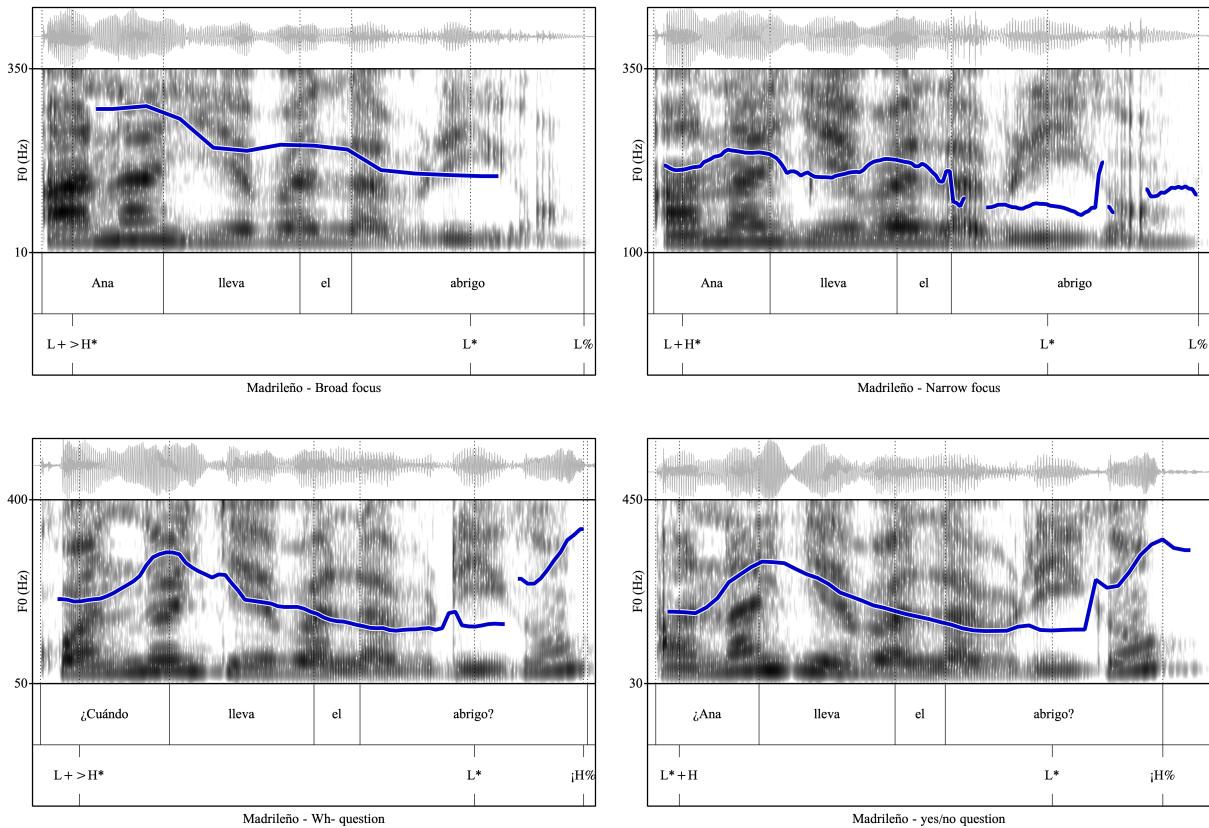


Figure 14. Waveform, spectrogram, and F0 trace exemplifying a broad focus (top left), narrow focus (top right), wh- question (bottom left), and yes/no question (bottom right) from a 29 year old female speaker of Madrileño Spanish.

Figure 15 plots the four utterance types as spoken by the Andalusian speaker. Similar to speaker of Madrileño Spanish, we do not see a major distinction between declaratives in nuclear position ($H+L^* L\%$). We found a tendency for this speaker to a use statement-of-the-obvious pattern where we elicited narrow focus. We show the nuclear configuration for wh- questions as $L+H^* L\%$, but this speaker also used $L+H^* L\%$. A final rise for wh- questions is also documented in this variety (Henriksen et al., 2010) and is associated with formality. For yes-no questions, the boundary tone was typically $\dot{H}\%$. More information on Western Andalusian intonation is available Sosa (1999) and Henriksen and García-Amaya (2012).

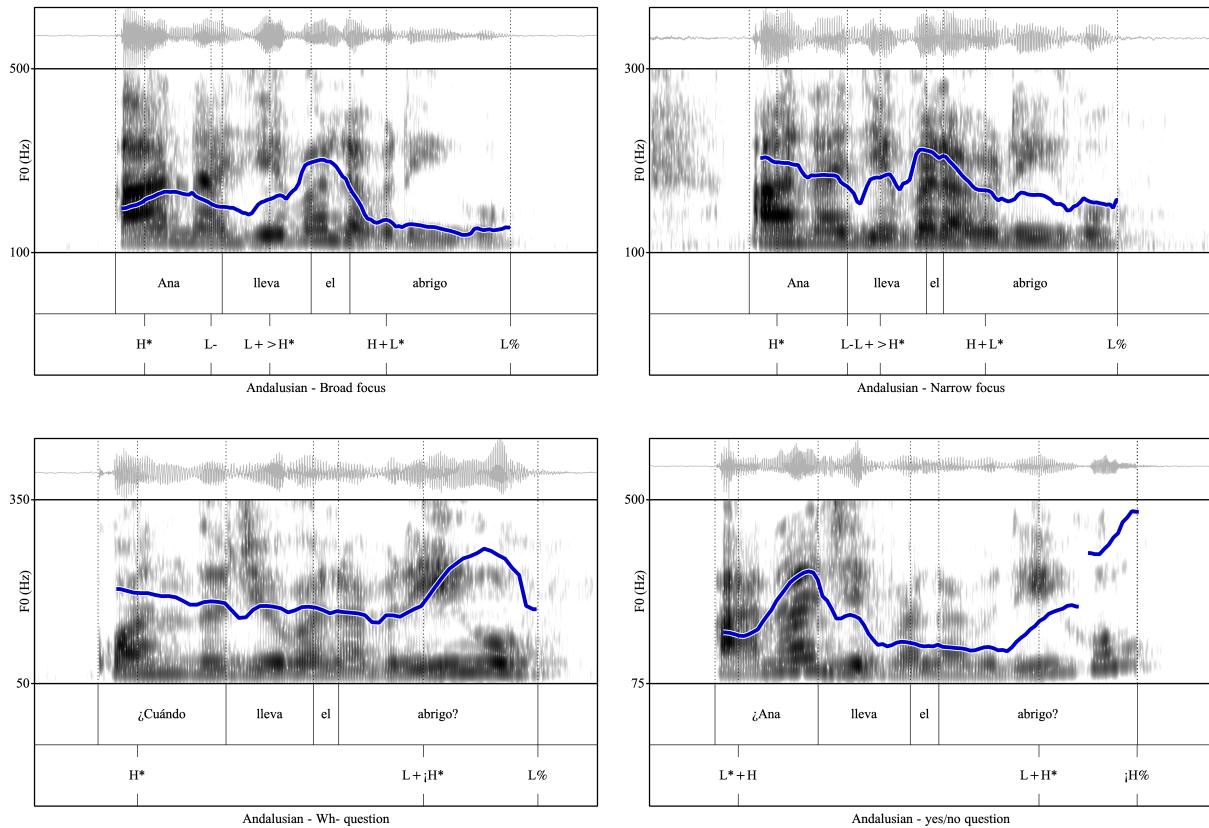


Figure 15. Waveform, spectrogram, and F0 trace exemplifying a broad focus (top left), narrow focus (top right), wh- question (bottom left), and yes/no question (bottom right) from a 35 year old female speaker of Andalusian (San Fernando, Cádiz) Spanish.

The Argentine speaker (Figure 16) generally produced declaratives with L* L% nuclear configurations. Notably absent from the narrow focus stimuli is the triton accent, L+H*+L (Labastía, 2006, among others; See Sosa, 1999). For wh- and yes/no questions, we see L% and HL% nuclear boundary tones, respectively. For more information on Argentine intonation we recommend Colantoni and Gurlekian (2004), Colantoni (2011), and Labastía (2011).

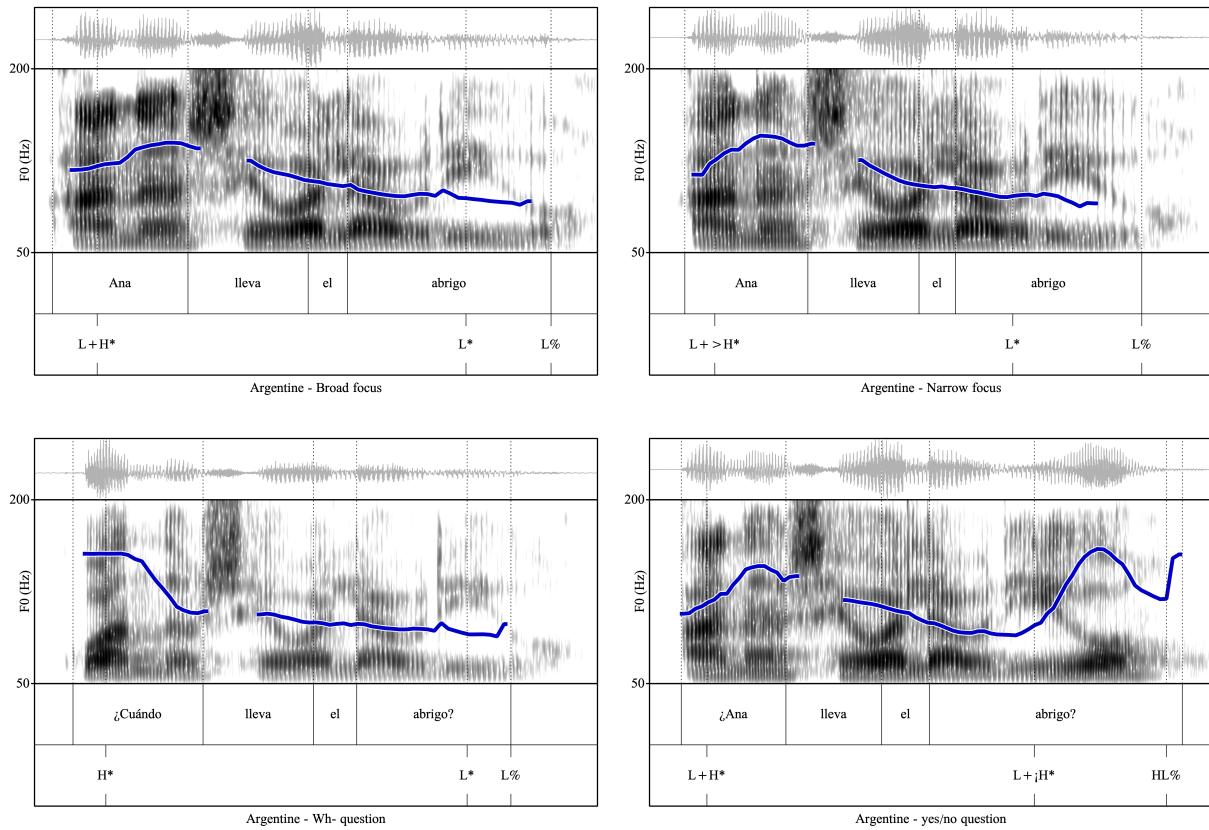


Figure 16. Waveform, spectrogram, and F0 trace exemplifying a broad focus (top left), narrow focus (top right), wh- question (bottom left), and yes/no question (bottom right) from a 27 year old male speaker of Argentine (Buenos Aires) Spanish.

Figure 17 illustrates example utterances from the Chilean speaker. We found that broad and narrow focus statements differed in their nuclear configuration ($H+L^* L\%$ vs. $L+H^* L\%$, respectively). We also noted that wh- and yes/no questions differed in the nuclear pitch accents (L^* vs. $H+L^*$, respectively). For more information on Chilean intonation, we refer the reader to Ortiz-Lira and Cid-Uribe (2000) and Ortiz-Lira (2003).

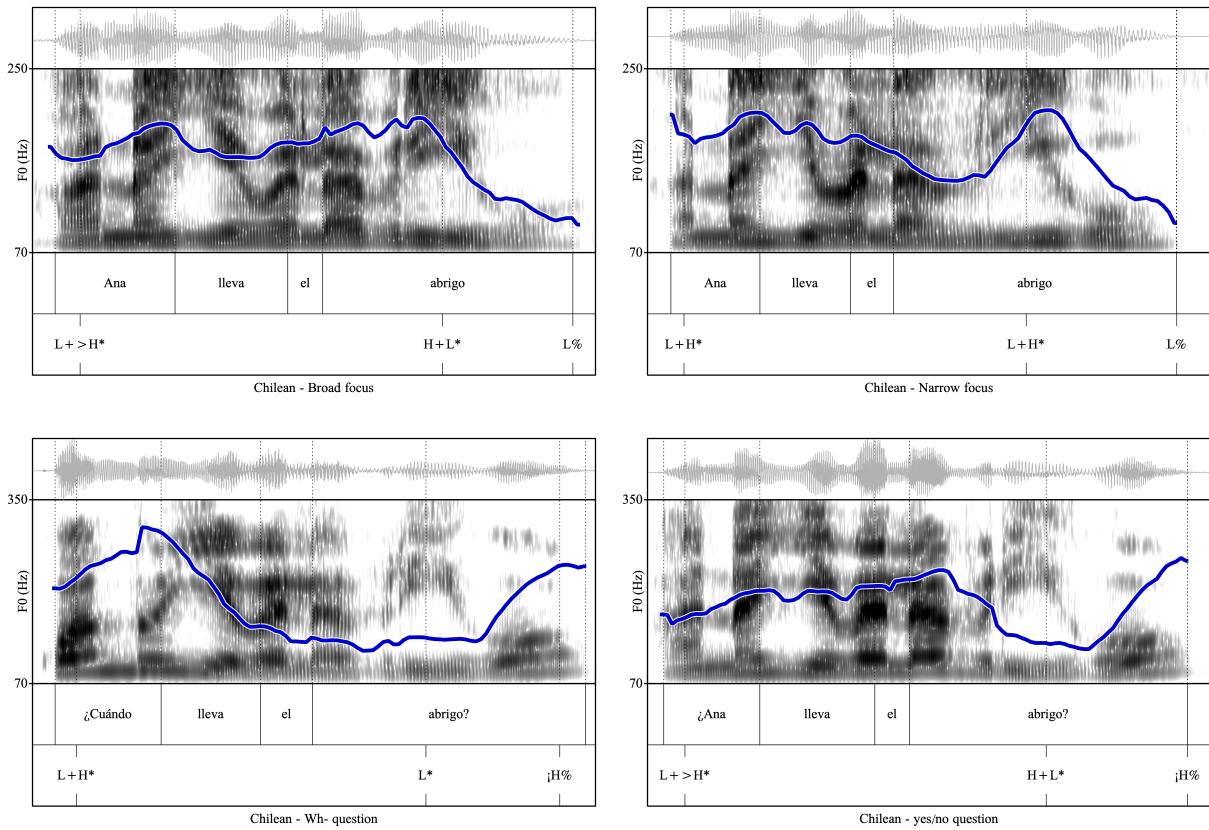


Figure 17. Waveform, spectrogram, and F0 trace exemplifying a broad focus (top left), narrow focus (top right), wh- question (bottom left), and yes/no question (bottom right) from a 42 year old female speaker of Chilean (Valparaíso) Spanish.

In Figure 18 we see example utterances from the Cuban speaker. We typically only found a distinction in broad and narrow focus items in the prenuclear pitch accents ($L+H^*$ and $L+>H^*$, respectively). Both questions showed followed falling patterns in nuclear position. For example, we found $L^* L\%$ in wh- questions and $L+H^* L\%$ in yes/no questions. In prenuclear position, our speaker also used $L+>H^*$ in yes/no questions. For more information, the interested reader can consult Sosa (1999), or Alvord (2006) for an account of Cuban Spanish in contact with English in Miami, Florida.

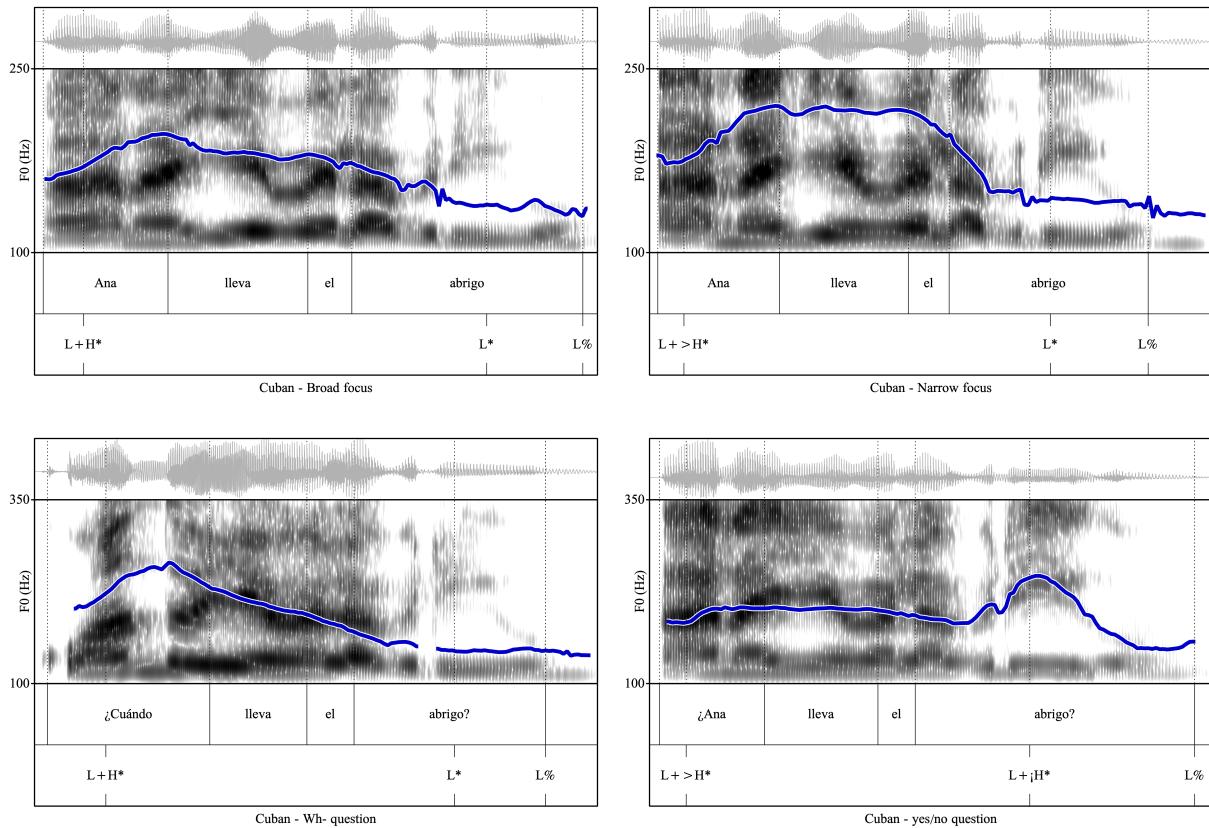


Figure 18. Waveform, spectrogram, and F0 trace exemplifying a broad focus (top left), narrow focus (top right), wh- question (bottom left), and yes/no question (bottom right) from a 55 year old female speaker of Cuban (Havana) Spanish.

The Mexican speaker distinguished between broad and narrow focus using a L* L% pattern in the former and L+H* H% in the latter. For wh- questions in this variety, final falls are considered the unmarked configuration (Sosa, 2003b), though we observed some rises, as in Figure 19 (bottom left plot). For yes/no questions this speaker produced a L* iH% nuclear configuration. Mexican Spanish is well-documented. We refer the reader to Quilis (1981), Quilis (1987), Quilis (1993), Sosa (1999), Butragueño (2003), and Butragueño (2004) for more information.

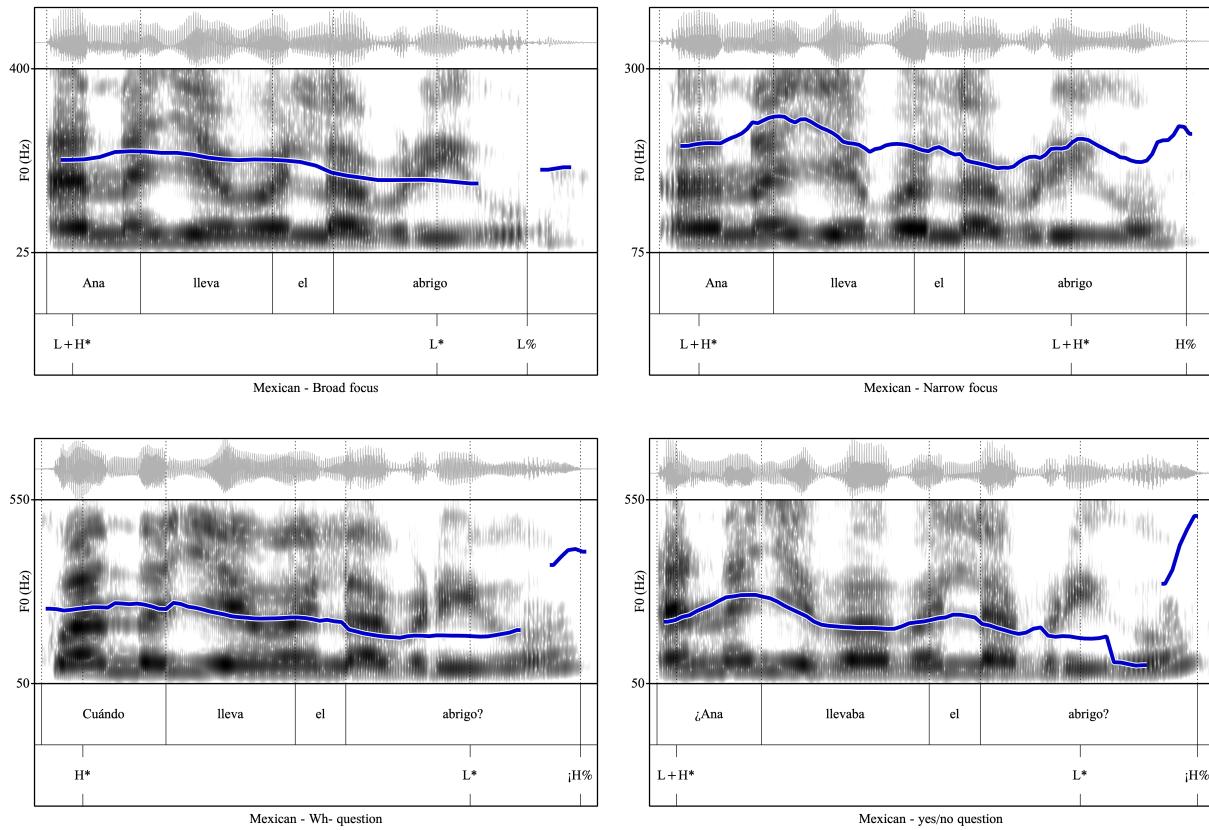


Figure 19. Waveform, spectrogram, and F0 trace exemplifying a broad focus (top left), narrow focus (top right), wh- question (bottom left), and yes/no question (bottom right) from a 30 year old female speaker of Mexican (Mexico City) Spanish.

The statement utterances from our speaker of Peruvian Spanish included a nuclear configuration of L* L%. The examples provided in the top row of Figure 20 are primarily differentiated by a prenuclear L+>H* in the broad focus condition and L+H* in the narrow focus condition. With regard to questions, this speaker often used a L+H* L% nuclear configuration for wh- questions, and L* iH% for yes-no questions. More information on Peruvian intonation is available in Sosa (1999) and O'Rourke (2005).

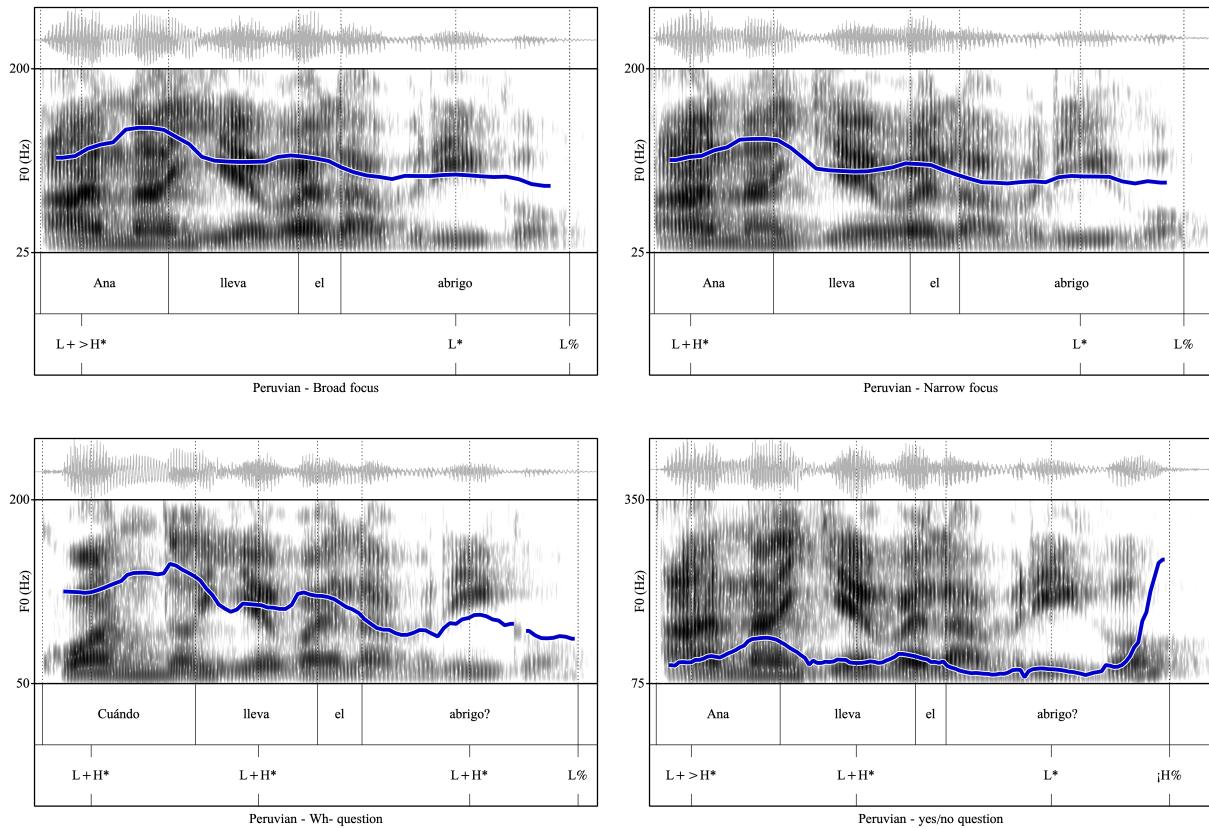


Figure 20. Waveform, spectrogram, and F0 trace exemplifying a broad focus (top left), narrow focus (top right), wh- question (bottom left), and yes/no question (bottom right) from a 30 year old male speaker of Peruvian (Lima) Spanish.

In Figure 21 we see example utterances from the Puerto Rican speaker. We typically only found a distinction in broad and narrow focus items in the prenuclear pitch accents ($L+H^*$ and H^* , respectively). On occasion this speaker produced final rises for wh- questions ($L^* \ jH\%$), but also often used a falling contour. For yes/no questions, this speaker produced $jH^* L\%$ configurations. More information on Puerto Rican intonation is available to the reader in Quilis (1981), Quilis (1987), Quilis (1993), Sosa (1999), Sosa (2003b), Armstrong (2010), and Armstrong (2012), among many others.

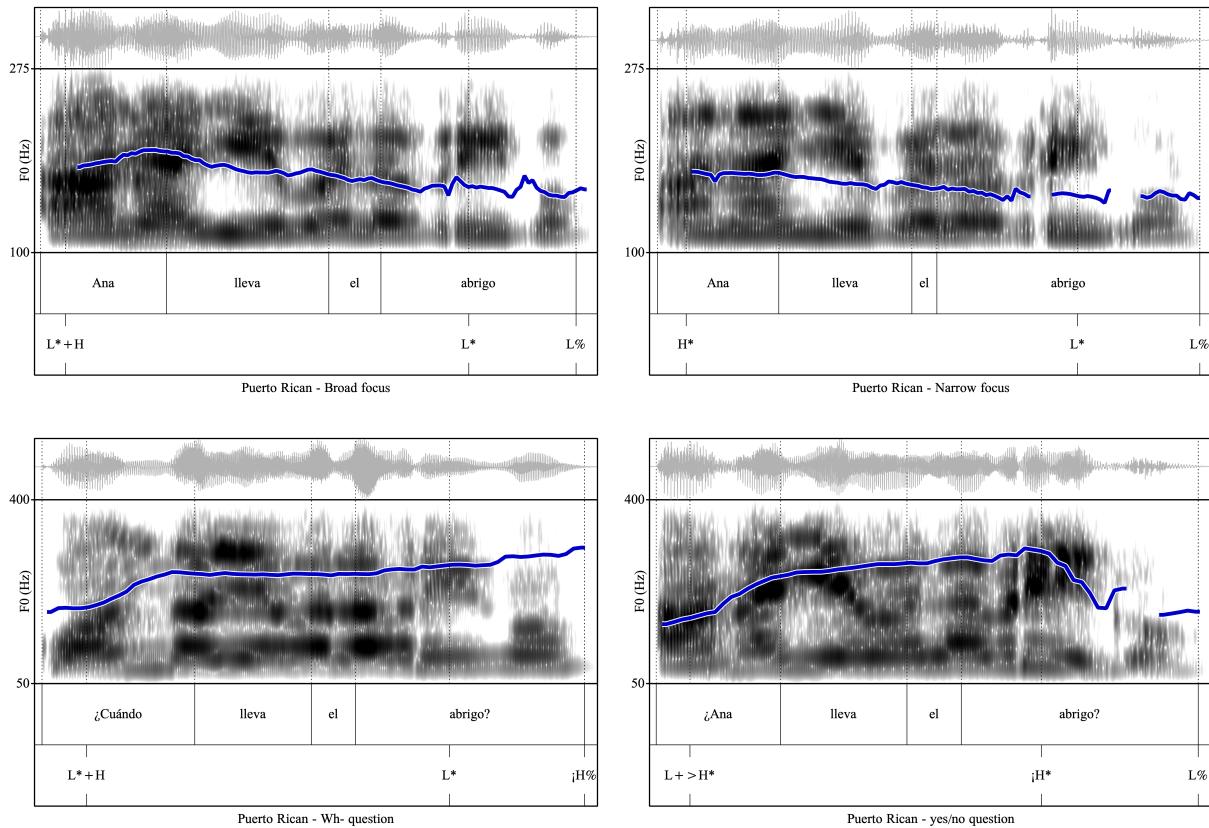


Figure 21. Waveform, spectrogram, and F0 trace exemplifying a broad focus (top left), narrow focus (top right), wh- question (bottom left), and yes/no question (bottom right) from a 35 year old female speaker of Puerto Rican (Ponce) Spanish.

In sum, we see that the acoustic stimuli used for the 2AFC task of the present project presents a large amount of between-variety and within-speaker variability. Most notably, our declarative utterances are often difficult to distinguish and we do not often see the typical $L+H^* L\%$ nuclear configuration that most commonly seen in the extant literature. We believe this is likely due to the fact that the manner in which we elicited narrow focus differs from other studies, which tend to use more specific pragmatic contexts (narrow focus correction, narrow focus contradiction), as opposed to a narrow focus utterance produced from answering an information seeking wh-question. That is, it may be the case that the use of a different pragmatic context and lab speech may result in contours different from those attested in other studies. This likely resulted in our narrow focus condition being easier than we initially intended. Our method of eliciting narrow focus follows that of Brandl et al. (2020), of which our study is a conceptual replication. We do not have access to their auditory stimuli, thus we cannot confirm which pitch contours were most common in their narrow focus condition. To illustrate the panhispanic and within-speaker variability of statements and questions, we plot the individual pitch contours from each utterance type by variety in Figure 22.

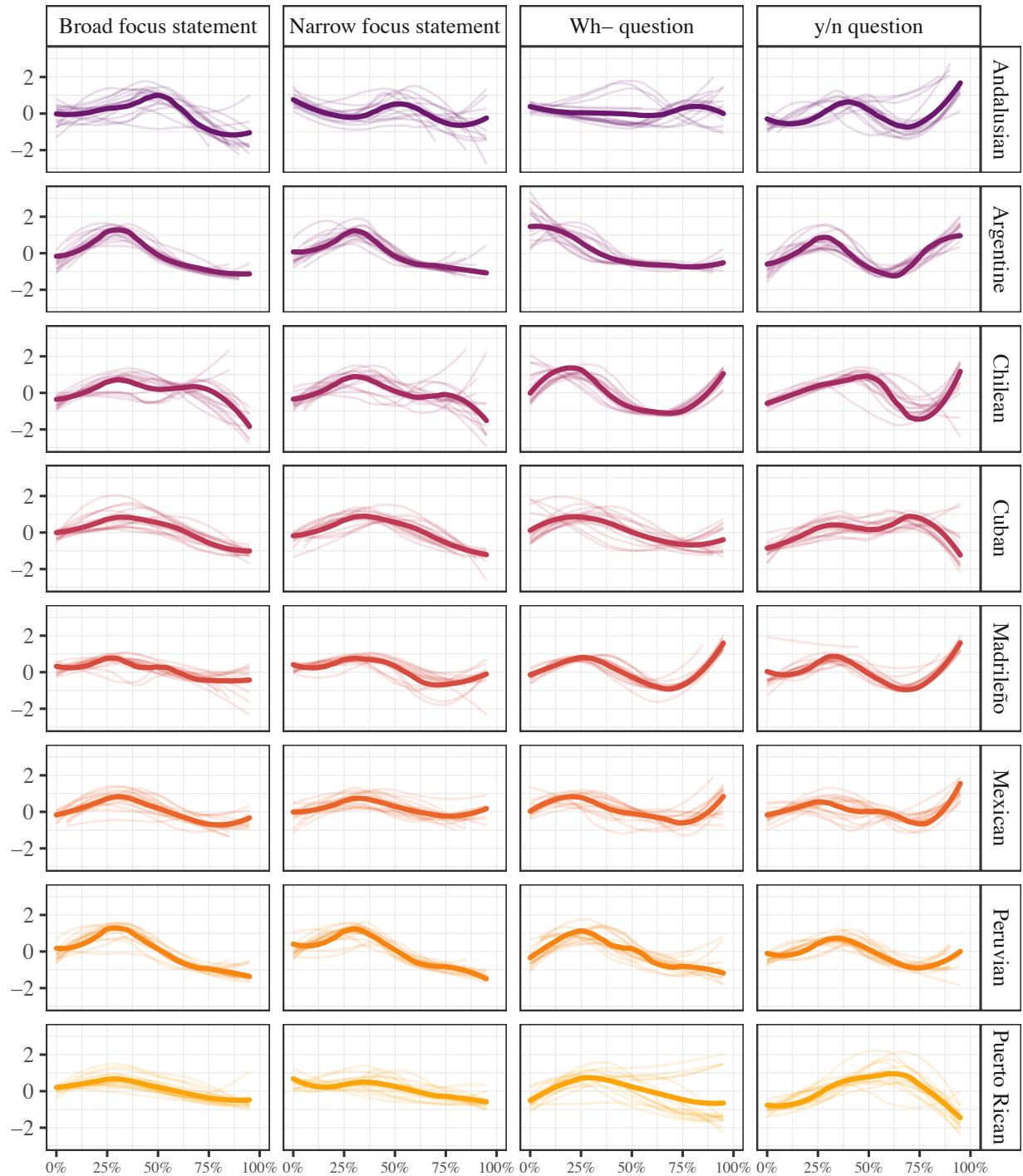


Figure 22. Pitch contours as function of time for utterance type and variety of Spanish. Thinner, opaque lines represent individual items, and thicker, dark lines represent the average contour trajectory. F0 values were log transformed and standardized for between speaker/variety comparability.

Randomization check across participants. For the purposes of our research questions, it was important that every participant be presented with stimuli from all of the Spanish varieties to which we had access. Recall that the 2AFC task contained 64 items, 16 of each utterance type. Using javascript we assigned each variety an equal probability of being selected in a given trial (0.125). To ensure that our randomization worked as planned (i.e., with each variety represented approximately equally across all trials and all participants), we calculated the average number of times each variety was presented in the data set ($n = 225$, and 14400 trials). One can observe in Figure 23 that this is indeed the case.

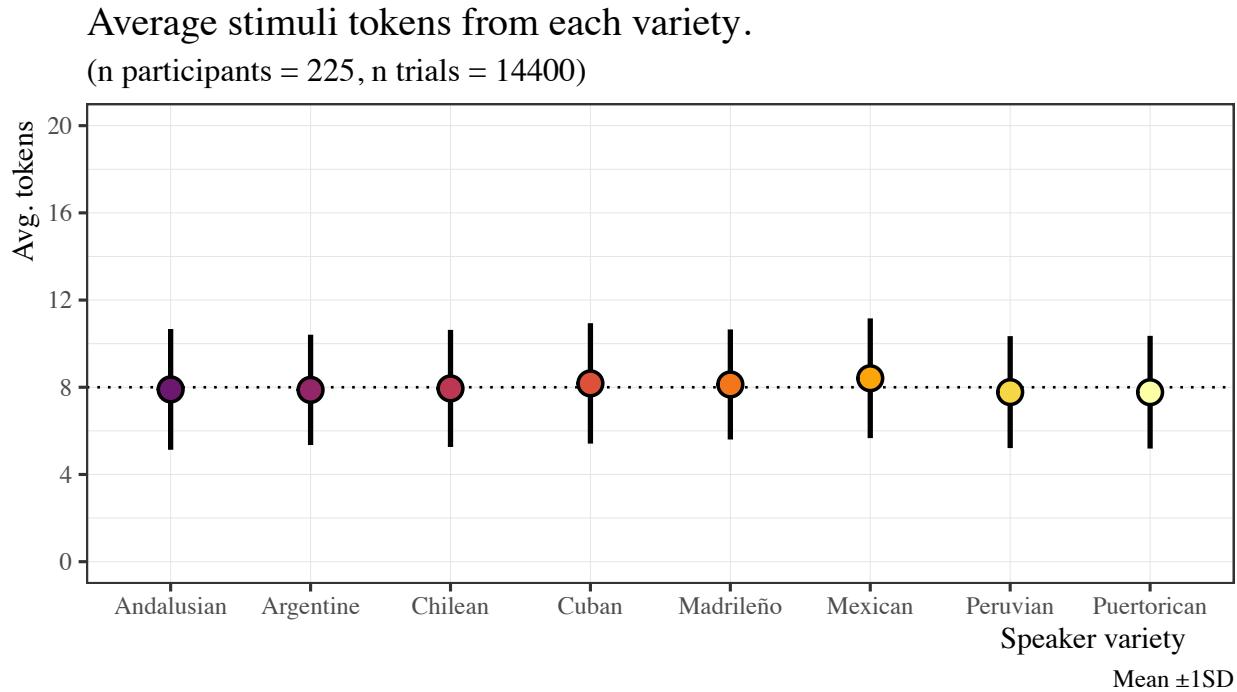


Figure 23. Average number of tokens (± 1 SD) presented from each speaker variety across all 14,400 trials. The experiment was programmed such that each of the 8 varieties had an equal probability of being presented (12.50%) across 64 experimental trials.

Items. Table 13 provides a list of all of the target items used for each utterance type.

Table 13: Experimental items produced in auditory stimuli.

Utterance type	Item
Broad focus statement	Ana lleva el abrigo
	Daniel iba a Bolivia
	David leía el libro
	El bebe comía muy bien
	El hombre mira la luna
	El niño oye el río
	Emilio ama la marcha
	La maestra vive en Paris
	La niña lava el plato
	Manuela vende el carro
	María bebe el vino

Utterance type	Item
	Mariano habla del tiempo Marta abre el regalo Mi madre come la fruta Mi novio viene del lago Mi tía odia la lluvia
Narrow focus statement	(¿Qué lleva Ana?) Ana lleva el abrigo (¿A dónde iba Daniel?) Daniel iba a Bolivia (¿Qué leía David?) David leía el libro (¿Cómo comía el bebé?) El bebe comía muy bien (¿Qué mira el hombre?) El hombre mira la luna (¿Qué oye el niño?) El niño oye el río (¿Qué ama Emilio?) Emilio ama la marcha (¿Dónde vive la maestra?) La maestra vive en Paris (¿Qué lava la niña?) La niña lava el plato (¿Qué vende Manuela?) Manuela vende el carro (¿Qué bebe María?) María bebe el vino (¿De qué habla Mariano?) Mariano habla del tiempo (¿Qué abre Marta?) Marta abre el regalo (¿Qué come tu madre?) Mi madre come la fruta (¿De dónde viene tu novio?) Mi novio viene del lago (¿Qué odia tu tía?) Mi tía odia la lluvia
Wh- question	¿Cuándo bebía el vino? ¿Cuándo comía la fruta? ¿Cuándo lavaba el plato? ¿Cuándo leía el libro? ¿Cuándo lleva el abrigo? ¿Cuándo miraba la luna? ¿Cuándo vendía el carro? ¿Por qué abre el regalo? ¿Por qué ama la navidad? ¿Por qué desayuna muy bien? ¿Por qué hablaba del agua? ¿Por qué iba a Bolivia? ¿Por qué odiaba la lluvia? ¿Por qué oia el río? ¿Por qué venia del lago? ¿Por qué vivia en Paris?
yes/no question	¿Ana lleva el abrigo? ¿Daniel iba a Bolivia? ¿David leía el libro? ¿El bebe comía muy bien? ¿El hombre mira la luna? ¿El niño oye el río? ¿Emilio ama la marcha? ¿La maestra vive en Paris? ¿La niña lava el plato? ¿Manuela vende el carro? ¿María bebe el vino? ¿Mariano habla del tiempo? ¿Marta abre el regalo? ¿Mi madre come la fruta?

Utterance type	Item
	<i>¿Mi novio viene del lago?</i>
	<i>¿Mi tía odia la lluvia?</i>

Author contributions

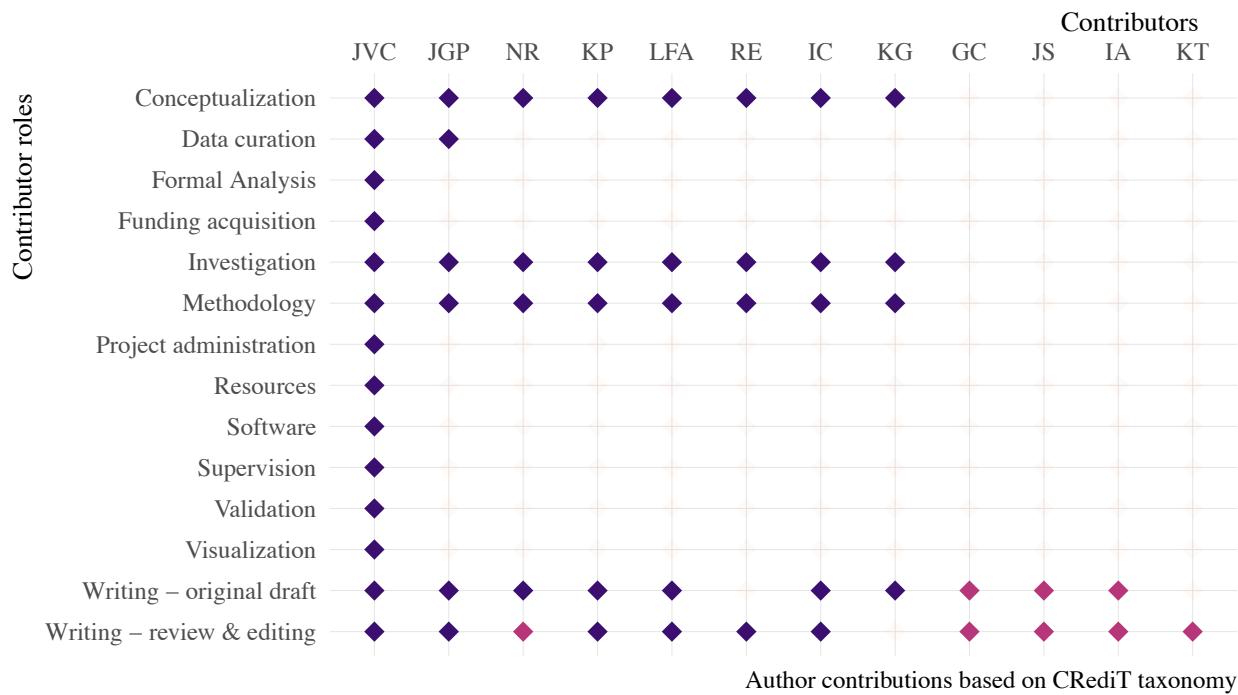


Figure 24. Author contributions according to the CREDiT author roles taxonomy. Contributions are indicated as being substantial (dark diamonds) or moderate (light diamonds).

Reproducibility information

About this document

This document was written in RMarkdown using `papaja` (Aust & Barth, 2018).

Session info

setting	value
version	R version 4.2.1 (2022-06-23)
os	macOS Big Sur ... 10.16
system	x86_64, darwin17.0
ui	X11
language	(EN)
collate	en_US.UTF-8
ctype	en_US.UTF-8
tz	America/New_York
date	2023-02-17
pandoc	2.19.2 @ /Applications/RStudio.app/Contents/Resources/app/quarto/bin/tools/ (via rmarkdown)
	loadedversion date
bayesplot	1.10.0 2022-11-16
bayestestR	0.13.0 2022-09-18
beeswarm	0.4.0 2021-06-01
bookdown	0.32 2023-01-17
brms	2.18.0 2022-09-19
cmdstanr	0.5.3 2022-08-25
coda	0.19-4 2020-09-30
contributor	0.3.0 2022-10-12
devtools	2.4.5 2022-10-11
dplyr	1.1.0 2023-01-29
effectsize	0.8.3 2023-01-28
emmeans	1.8.4-1 2023-01-17
forcats	1.0.0 2023-01-29
fs	1.6.0 2023-01-23
ganttrify	0.0.0.9008 2022-10-12
ggbeeswarm	0.7.1 2022-12-16
ggdist	3.2.1 2023-01-18
ggplot2	3.4.0 2022-11-04
ggrepel	0.9.3 2023-02-03
ggridges	0.5.4 2022-09-26
glue	1.6.2 2022-02-24
here	1.0.1 2020-12-13
insight	0.19.0 2023-01-30
knitr	1.42 2023-01-25
loo	2.5.1 2022-03-24
magrittr	2.0.3 2022-03-30
mvttnorm	1.1-3 2021-10-08
nlme	3.1-162 2023-01-31
papaja	0.1.1 2022-07-05
parameters	0.20.2 2023-01-27
patchwork	1.1.2 2022-08-19

posterior	1.3.1	2022-09-06
prettyunits	1.1.1	2020-01-24
printy	0.0.0.9003	2022-10-17
purrr	1.0.1	2023-01-10
Rcpp	1.0.10	2023-01-22
RcppParallel	5.1.6	2023-01-09
readr	2.1.3	2022-10-01
readxl	1.4.1	2022-08-17
remotes	2.4.2	2021-11-30
rlang	1.0.6	2022-09-24
rmarkdown	2.20	2023-01-19
rprojroot	2.0.3	2022-04-02
rstan	2.21.8	2023-01-17
rstantools	2.2.0	2022-04-08
scales	1.2.1	2022-08-20
sessioninfo	1.2.2	2021-12-06
stringr	1.5.0	2022-12-02
tibble	3.1.8	2022-07-22
tidybayes	3.0.3	2023-02-04
tidyverse	1.3.0	2023-01-24