Affective Science (2021) 2:301–310
https://doi.org/10.1007/s42761-021-00038-w

RESEARCH ARTICLE

Categorization of Vocal Emotion Cues Depends on Distributions of Input

Kristina Woodard[1] · Rista C. Plate[1,2] · Michele Morningstar[3] · Adrienne Wood[4] · Seth D. Pollak[1]

Received: 27 July 2020 / Accepted: 9 February 2021 / Published online: 10 April 2021

Abstract

Learners use the distributional properties of stimuli to identify environmentally relevant categories in a range of perceptual domains, including words, shapes, faces, and colors. We examined whether similar processes may also operate on affective information conveyed through the voice. In Experiment 1, we tested how adults (18–22-year-olds) and children (8–10-year-olds) categorized affective states communicated by vocalizations varying continuously from "calm" to "upset." We found that the threshold for categorizing both verbal (i.e., spoken word) and nonverbal (i.e., a yell) vocalizations as "upset" depended on the statistical distribution of the stimuli participants encountered. In Experiment 2, we replicated and extended these findings in adults using vocalizations that conveyed multiple negative affect states. These results suggest perceivers' flexibly and rapidly update their interpretation of affective vocal cues based upon context.

Keywords Emotion categorization · Vocal expression · Affect · Statistical learning

Central theoretical questions in emotion research concern the issue of learnability. Emotion cues vary, to some extent, across individuals, situations, groups, and cultures. This variability may be construed as evidence that emotions would be too difficult to learn without the existence of some stable emotion categories across individuals or, conversely, that the amount of variability itself argues against any core set of emotion categories. Both of these positions have been articulated in the literature (Barrett et al., 2019; Keltner et al., 2019; Scarantino, 2014). Yet, there is no question about the fact that despite the variability encountered, children and adults continue to refine and use perceptual categories to systematically distinguish between emotional states. The present study explores whether the perceptual variability in emotion cues is itself an important source of emotion learning. To do so, we examine whether individuals can track and use variability in the distribution of the cues they encounter in their perceptual categorization of emotion.

One path to progress on this question about the role of variability in emotion learning is to draw from advances in the field of language learning. A central question in the field of speech perception has been how to understand the efficiency of speech perception given the "lack of invariance" in the input people receive (Liberman, 1957; Liberman et al., 1967). A phoneme—the smallest unit of sound that makes up language, like the /ɛ/ in "pen"—can sound very different depending on who is saying it, the context or co-articulation of the phoneme with other aspects of an utterance, speech conditions, and random errors in speech production (Miller & Eimas, 1995). As an example, the /ɛ/ in "pen" will sound slightly different depending on whether it is produced by an adult or child, how fast or slow someone is speaking, the dialect of the speaker, and the sounds produced right before or after the word (e.g., Kleinschmidt & Jaeger, 2011). Yet despite this variation, humans learn to perceive speech sounds categorically, quickly, and with high accuracy.

Handling Editor: Disa Sauter

✉ Kristina Woodard
kwoodard2@wisc.edu

[1] Department of Psychology, University of Wisconsin – Madison, 1500 Highland Avenue, Madison, WI 53705, USA

[2] Department of Psychology, University of Pennsylvania, 3720 Walnut Street, Philadelphia, PA 19104, USA

[3] Department of Psychology, Queen's University, Kingston, ON, Canada

[4] Department of Psychology, University of Virginia, 485 McCormick Rd, Charlottesville, VA 22904, USA

One mechanism that supports perceptual learning of speech cues is humans' ability to track and adapt to the distributional properties of acoustic input (Kleinschmidt & Jaeger, 2015; Samuel & Kraljic, 2009). Categorization of phonemes is surprisingly relative as well as context- and speaker-dependent. For example, individuals are able to quickly adjust to different rates of speaking (Newman & Sawusch, 1996), differences in speech caused by vocal tract size (Johnson, 2005), foreign-accented speech (Clarke & Garrett, 2004; Xie et al., 2018), dialects (Dahan et al., 2008), and variation in vowel pronunciation (Weatherholtz, 2015). These rapid adaptations are not always temporary; rather, they can update perceivers' representations over time and transfer to novel situations and speakers (Clarke-Davidson et al., 2008; Kleinschmidt, 2019; Weatherholtz & Jaeger, 2016; Xie et al., 2018).

Here, we examine whether learning that is based upon the distributional properties of perceptual input also applies to vocal emotion cues. This type of learning has already been implicated in a range of developmental processes that includes children's category learning in language (Saffran, 2020), faces (Dotsch et al., 2017), color, and action sequences (see Frost et al., 2019 for review). For instance, distributional statistics can aid children's language learning by allowing them to detect phoneme categories (Maye et al., 2002) and can influence adults' color perceptions based on the amounts of each color in their current environment (Levari et al., 2018). Recent findings suggest that distributional information also influences the learning of visual facial cues for emotion categories (Levari et al., 2018; Plate et al., 2019; Plate et al., in press). Brief exposure to images of a person who was either facially expressive or unexpressive caused children and adults to shift their threshold for categorizing a face as emotional.

The vocal expression of affect parallels these other domains in many respects. Specifically, there are statistical consistencies in how some affective states are conveyed (Banse & Scherer, 1996; Juslin & Laukka, 2001; Sauter et al., 2010). For example, anger is often conveyed with high pitch, high intensity, and, if using spoken word, rapid speech rate (Johnstone & Scherer, 2000; Scherer, 2019). However, vocal affect also reflects a "lack of invariance": similar vocal properties (such as high pitch) can predict different emotions, speaker differences that affect speech perception can also impact vocal emotion, and there are currently no one-to-one mappings between any combinations of acoustic features and a specific emotion (Ito, 2018; Sauter et al., 2010).

The present work examines whether similar distributional learning processes also operate on affective information conveyed through the voice. Testing this idea is important because some reports suggest that this type of learning may be specific to some modalities, such as auditory versus visual, or specific to certain kinds of stimuli, evidenced by the lack of transfer learning of novel stimuli (Frost et al., 2015). Moreover, learning performance across different modalities is often weakly correlated (Siegelman & Frost, 2015), making it important to test assumptions of generalizability. We also examined this learning process in children, who are still acquiring vocal emotion categories, based upon data indicating that prior knowledge and experience with stimuli affects statistical learning (Siegelman et al., 2018). To do so, we tested 8- to 10-year-olds because children at this age can rely on either lexical or prosodic information to interpret auditory expressions of emotion (e.g., Friend & Bryant, 2000; Morton & Trehub, 2001), but have lower accuracy than adults when identifying auditory emotion categories (Aguert et al., 2013; Morningstar, Ly, Feldman, & Dirks, 2018; Morningstar, Nelson, & Dirks, 2018).

We tested how perceivers categorized nonverbal (i.e., a yell) and verbal (i.e., spoken word, with hostile tone) auditory stimuli of different emotional intensities as either "calm" or "upset." We include both verbal and nonverbal stimuli as both adults and children tend to have higher accuracy recognizing nonverbal vocalizations (Hawk et al., 2009; Sauter et al., 2013), and this could impact how individuals adjust to these vocalizations. In Experiment 1, children and adults were trained to a baseline and then exposed to different distributions of vocal stimuli—that is, vocal stimuli with different ranges of intensity. Thus, after training, some participants were exposed to a greater proportion of vocal cues at higher intensities (upset shifted), some participants were exposed to a greater proportion of vocal cues at lower intensities (calm shifted), and some participants were exposed to the same proportion of stimuli throughout the entire study (unshifted).

We predicted that adults and children exposed to these different ranges of vocal stimuli would adjust how they categorized whether or not vocalizations were "upset." For instance, participants who were exposed to more intense ranges might categorize certain vocalizations as "calm," whereas participants exposed to less intense ranges might categorize this same stimulus as "upset." We predicted these changes in categorization because the distributions of stimuli encountered are giving different information about how expressive the individual is. Comparing the performance of adults with children, who are still acquiring emotion categories, afforded the opportunity to examine developmental differences in how representations of affective vocal cues are updated. If children, in addition to adults, exhibit such sensitivity to statistical distributions of vocalizations, then statistical learning might support initial acquisition of emotion cue categories just as it does learning in other domains. In Experiment 2, we use the same paradigm to test the replicability of this perceptual mechanism and determine if the effects hold when individuals also need to track other negatively valenced emotions and speakers.

Experiment 1

Method

Participants

Eighty-four children (41 female; age range = 8–10 years, Mage = 9.70 years, SDage = 0.88 years) and 87 adults (58 female; Mage = 19.10 years, SDage = 0.73 years) participated in this experiment. We had three between-subject conditions and aimed for 30 participants in each condition (90 total) based on sample sizes of previous research (Experiment 1 in Plate et al., 2019), however, we ended up slightly short of our recruitment goal because of the COVID-19 outbreak. Since we ended data collection early, we report post hoc power analyses with the results; these should be interpreted with the understanding that they were conducted after data collection (see Zhang et al., 2019). Two children completed only one condition (verbal). Children were recruited from the community in Madison, Wisconsin (8.33% African American, 7.14% Asian American, 4.76% Hispanic, 2.38% more than one race, 77.38% White). All children received a prize, and parents received $25 for their participation. Adults were undergraduate students at the University of Wisconsin-Madison and received course credit (2.30% African American, 19.54% Asian American, 11.49% Hispanic, 5.75% more than one race, 60.92% White). The Institutional Review Board approved the research.

Stimuli

We presented participants with both nonverbal and verbal auditory stimuli. Nonverbal stimuli, created using Soundgen, were based on a male vocalization of a "yell" (roar_059; Anikin & Persson, 2017). Validation for the stimuli is reported in Anikin (2019) and R scripts generating the stimuli and the stimuli themselves are available at: https://osf.io/749xq/?view_only=ef7f9d9509284ed2927948509c596db5). Twenty-one nonverbal morphs were generated. These stimuli were morphs from a neutral "ahh" (0% "upset") to a hostile "ahh" (100% "upset") that varied in 5% increments in features including pitch, amplitude/loudness, and other cues of vocal quality (Banse & Scherer, 1996; Juslin & Laukka, 2001; Sauter et al., 2010). Verbal stimuli were morphs of recordings of a male actor saying a statement ("I can't believe you just did that") in both a neutral voice and an angry voice (see Morningstar et al., 2017 for details about recording procedure). Morphs were created by linearly manipulating the waveform in 10% increments of the actor's original portrayals from neutral to hostile, using STRAIGHT acoustic manipulation tools (Kawahara et al., 2008) in Matlab. The STRAIGHT tool manipulates F0, amplitude, spectral envelope, and periodicity simultaneously at the spectrogram level (Kawahara & Morise, 2011). This procedure yielded 10 recordings ranging from 10% emotional intensity (i.e., 10% anger, 90% neutral) to 100% emotional intensity (i.e., 100% anger, 0% neutral). Validation of these stimuli in a forced-choice emotion recognition task suggests that listeners' (n = 190) capacity to identify these recordings as "angry" increased with the emotional intensity of the morphs (Morningstar et al., under review), going from 11% accuracy for the 10% intensity recording to 86% accuracy for the 100% intensity recording (where chance was 14%). The difference in morphing increments for the verbal and nonverbal conditions occurred because of the stimuli availability. Additional details about the creation and validation of stimuli are available in the Supplemental Material.

Procedure

The present task tested how perceivers categorized auditory cues of varying intensity as either "calm" or "upset." The experiment included three phases: (1) a practice phase, (2) a training phase, and (3) a testing phase. The training phase gave participants explicit feedback on whether each cue should be categorized as "upset" or "calm" in order to create a baseline category boundary. The testing phase examined whether the category boundary established in the training phase would shift in response to different statistical distributions of stimuli (e.g., in response to hearing more or less upset vocalizations) in one of three conditions: calm shifted, unshifted, or upset shifted. Participants completed the entire procedure (practice, training, and testing) for both verbal and nonverbal stimuli, with order counterbalanced across participants such that half participated first in the verbal condition and half participated first in the nonverbal condition. Stimuli were presented with PsychoPy (v1.83.04).

Practice Phase During the practice phase, participants were introduced to "John" (neutral image of Actor 24 from the MacArthur Network Face Stimuli Set; Tottenham et al., 2009) and told that, "Just like everyone, sometimes John feels upset and sometimes he feels calm. Today we need your help figuring out if he is upset or calm." Participants were then taught that when John is feeling upset, he likes to "go to the red room and practice boxing," and when he is feeling calm, he likes to "go to the blue room and read a book." The goal of this design was to task participants with predicting the next action of the speaker based on their vocalization. On each trial, participants saw an image of headphones, and had to click on the headphones when they were ready to hear John make a sound. After hearing the sound, participants selected either a red room with an image of a punching bag (indicating they think he feels "upset") or a blue room with an image of an easy chair and book (indicating they think he feels "calm") using a computer mouse (see Supplemental Material, Figure S1). The

side of the screen where each room appeared was counterbalanced between participants. Participants completed 6 practice trials with feedback ("Correct!" or "Incorrect! Please try again") and repeated incorrect trials until they responded correctly. These practice trials included three calm trials (0%, 10%, 20% upset morphs were labeled as "calm") and three upset trials (80%, 90%, and 100% upset morphs were labeled as "upset"). The order of morphs was randomized.

Training Phase During the training phase, participants completed 24 trials with feedback in random order. Stimuli consisted of morphs ranging from 20% upset to 80% upset. The 50% morph was omitted in order to emphasize the category boundary at the midpoint. Participants received feedback ("Correct!" or "Incorrect!") after each trial, with "upset" being the correct response for morphs greater than 50% upset, and "calm" being the correct response for morphs less than 50% upset.

Testing Phase During the testing phase participants completed 72 trials in random order. Participants were randomly assigned to one of three conditions: calm shifted, unshifted, and upset shifted. In the unshifted condition, participants heard the same stimuli as in the training phase (20% upset to 80% upset with the 50% morph omitted to create a category boundary). In the upset shifted condition, participants heard stimuli with a higher average percentage of intensity (40% upset to 100% upset with the 70% morph omitted to create a category boundary). In the calm shifted condition, participants heard stimuli with a lower average percentage of intensity (0% upset to 60% upset with the 30% morph omitted to create a category boundary). No feedback was given to participants during this phase.

Results

We sought to analyze whether adults and children flexibly shifted their category boundaries—the point on a morph continuum where they switched from categorizing the stimuli as "calm" to "upset"—for both verbal and nonverbal vocalizations based upon the distributional sampling of the stimuli they encountered. First, we evaluated whether participants were able to learn the category boundary during the training phase, and whether adults and children were able to similarly learn this boundary. Determining participant behavior during training ensures that differences observed at testing resulted from the distributions of the stimuli, rather than some feature of the stimuli or response biases that participants had prior to participation in the experiment. Next, we evaluated if participants' category boundaries changed based upon their exposure to the distribution of stimuli during the training phase. Figures depicting the training phase performance are available in the Supplemental Material. We analyzed verbal and nonverbal conditions separately because our hypotheses were formulated around learning based upon probabilistic sampling of perceptual input, and we did not have a priori hypotheses about differences across stimuli. However, we do present a post hoc comparison of the verbal and nonverbal conditions in the Supplemental Materials. Analyses were completed in R version 3.6.2 (R Core Team, 2019) using the tidyverse package (Wickham et al., 2019), the lme4 package for our mixed-effects models (Bates et al., 2015), the ggplot2 (Wickham, 2016), and sjPlot (Lüdecke, 2020) packages for our graphs and tables, and the simr package for power analyses (Green & MacLeod, 2016). Stimuli, data, task scripts, and R scripts are available online at https://osf.io/749xq/?view_only=ef7f9d9509284ed2927948509c596db5. Although the participants who produced these stimuli did not consent to sharing the stimuli publicly online, they did give us permission to share the stimuli with other researchers upon request.

Do Perceivers Shift Their Categorization of Verbal Vocalizations Based upon the Distribution of Stimuli They Encounter?

Training Phase Both adults and children had high accuracy (children mean accuracy = 97.4%; adult mean accuracy = 96.6%), where accurate means labeling a sound more than 50% upset as "upset" (by clicking the red room) and less than 50% upset as "calm" (by clicking the blue room). To look at whether or not there were any group differences in accuracy, we ran a logistic generalized linear mixed-effects models predicting accuracy based on Age (children coded as − 0.5 and adults coded as 0.5) with a by-participant random intercept. We found no difference in accuracy between adults and children, $b = -0.24$, $z = -1.13$, $p = .26$, OR = 0.78. We also tested for group differences in how likely adults and children were to categorize each morph as "upset" (whether one age group was more likely to identify morphs as upset earlier or later in the continuum). To examine this, we used logistic generalized linear mixed-effects models on children's categorization of the vocal expressions ("calm" = 0, "upset" = 1) with a main effect of the Percent Upset of the stimuli (0% to 100% upset, mean-centered in increments of 10%), a main effect of Age (Children vs. Adults), the interaction between Percent Upset and Age, a by-participant random slope for Percent Upset, and a by-participant random intercept. Overall, no age differences in performance emerged during the training phase. Both adults and children similarly learned the category boundary, with vocal stimuli being more likely to be categorized as "upset" with each 10% increase in intensity, $b = 2.79$, $z = 17.68$, $p < .001$, OR = 16.28. There were no age-related differences in learning the category boundary, $b = -0.14$, $z = -0.65$, $p = .52$, OR = 0.87, and no interaction

between Age and Percent Upset, $b = -0.21$, $z = -1.02$, $p = .31$, OR = 0.81.

Testing Phase Next we examined whether participants would shift their emotion category boundaries after unsupervised exposure to a new statistical distribution of vocal input without feedback. We again used a logistic generalized linear mixed-effects model. The full model regressed participant responses on a three-way interaction between Percent Upset (mean-centered), dummy-coded Shift Type (calm shifted, unshifted, upset shifted), and Age (Children vs. Adults) plus all lower-order fixed effects, and a by-participant random slope for Percent Upset and a by-participant random intercept.

As in the training phase, there were no differences in performance between adults and children, $b = 0.07$, $\chi^2(1) = 0.03$, $p = 0.86$, OR = 1.07. As predicted, exposure to shifted distributions of vocal stimuli affected participants' categorization, $\chi^2(2) = 489.07$, $p < .001$ (Fig. 1). Those in the calm shifted condition identified vocal stimuli as upset earlier in the morph continuum, $b = 3.09$, $z = 10.86$, $p < .001$, OR = 21.97, while those in the upset shifted condition identified vocal stimuli as upset later in the morph continuum, $b = -3.19$, $z = -10.84$, $p < .001$, OR = 0.04. Those in the unshifted condition also had a steeper category boundary between identifying stimuli as "upset" versus "calm," which was not unexpected as individuals in this condition did not have to learn a different category boundary from training, $\chi^2(2) = 21.66$, $p < .001$ (interaction between Percent Upset and Shift Type, both dummy-coded terms were significant in the expected direction as well). These results indicate that participants adapted their categories about which auditory cues constituted "upset" based on the distribution of auditory morphs encountered in the shifted experimental conditions.

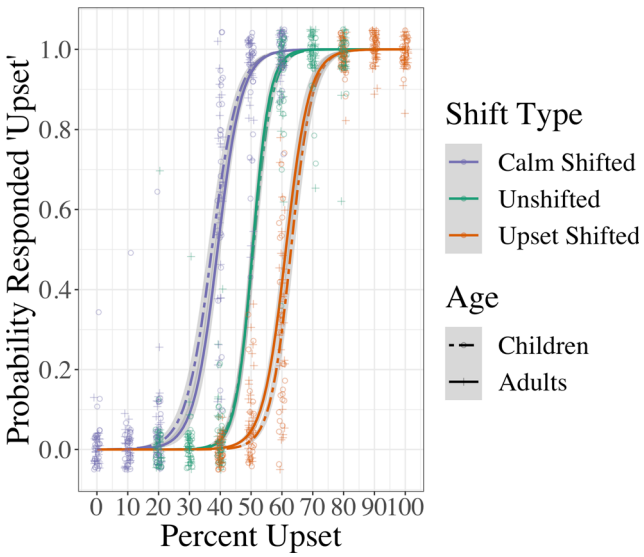

Fig. 1 Verbal testing phase: exposure to varying distributions of verbal stimuli affected participant's categorization

We conducted a post hoc power analysis by running 100 simulations in the SIMR package (Green & MacLeod, 2016) and found that we had essentially 100% power (95% CI: 96.36–100%) to detect our effect of Shift Type and 99% power (95% CI: 94.55–99.97%) to detect the Shift Type * Percent Upset interaction.

Do Perceivers Shift Their Categorization of Nonverbal Vocalizations Based upon the Distribution of Stimuli They Encounter?

Training Phase We examined participants' categorization of nonverbal vocalizations, using the same analytic models. Adults and children learned the emotion category boundary during training (children mean accuracy = 87.3%; adult mean accuracy = 90.3%); adults were slightly more accurate than children, $b = 0.31$, $z = 2.81$, $p < .01$, OR = 1.36. Next, we examined if there were differences in how likely adults and children were to categorize each morph as "upset." We again found that adults and children were able to learn the category boundary, with auditory stimuli being more likely to be categorized as "upset" with each 10% increase in intensity, $b = 1.66$, $z = 25.74$, $p < .001$, OR = 5.24. There was no main effect of Age, $b = 0.11$, $z = 0.96$, $p = 0.34$, OR = 1.12, indicating that children were not identifying morphs as upset earlier or later in the morph continuum than adults. However, there was an interaction between Age and Percent Upset, $b = 0.29$, $z = 2.66$, $p < .01$, OR = 1.33, indicating that adults had a steeper category boundary between "calm" and "upset" than children (reflecting the children's slightly higher error rate). These results indicate that both adults and children successfully learned the 50% category boundary during the training phase, but that children made more errors and had less precise category boundaries.

Testing Phase Next we examined whether exposure to a new statistical distribution of auditory input would shift participants' categorization. We used the same logistic generalized linear mixed-effects model as above. Unsupervised exposure to shifted distributions of nonverbal auditory stimuli again impacted participants' categorization behaviors, $\chi^2(2) = 707.84$, $p < .001$ (Fig. 2). Those in the calm shifted condition identified vocal stimuli as "upset" earlier in the morph continuum, $b = 3.01$, $z = 12.98$, $p < .001$, OR = 20.27, while those in the upset shifted condition identified vocal stimuli as "upset" later in the morph continuum, $b = -3.93$, $z = -16.65$, $p < .001$, OR = 0.02. There was no main effect of Age, $b = 0.36$, $\chi^2(1) = 1.53$, $p = 0.22$, OR = 1.44, no age-related interactions, and no significant interactions in the model. These data indicate that participants adapted their categories about which auditory cues constituted "upset" based on the distribution of auditory morphs they encountered in the experimental conditions.

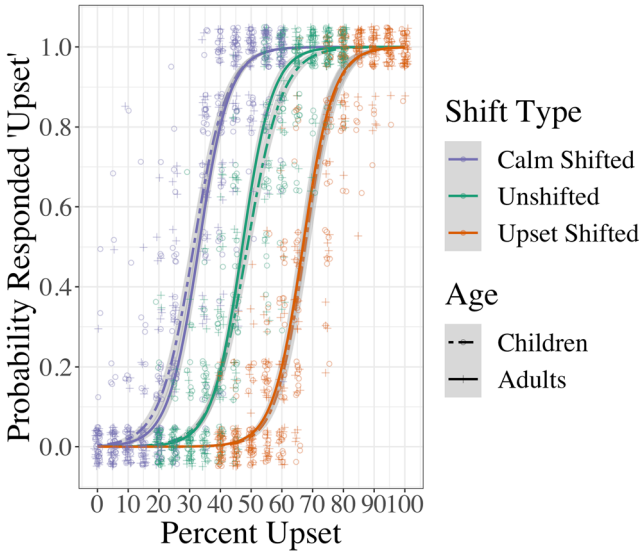

Fig. 2 Nonverbal testing phase: exposure to varying distributions of nonverbal auditory stimuli affected participant's categorization

We conducted a post hoc power analysis by running 100 simulations in the SIMR package (Green & MacLeod, 2016) and found that we had essentially 100% power (95% CI: 96.38–100.00) to detect our effect of Shift Type.

Discussion of Experiment 1

Participants shifted their categorization of both verbal and nonverbal auditory stimuli based on the distributions of input they encountered. Adults and children learned and adapted to the variation in input with similar flexibility and speed, except that children found the training phase of the nonverbal task slightly more difficult (see Supplemental Material for more detail).

Experiment 2

Experiment 2 tests whether the learning effects observed in Experiment 1 continue to emerge beyond the presentation of just a single prototypical cue. To extend the findings of Experiment 1, we used the same general procedure, but with a few key changes. First, we used multiple emotion categories in Experiment 2, which allowed us to examine the effect of Shift Type (calm shifted, unshifted, and upset shifted) as a within-participant manipulation, with each emotion assigned to a different Shift Type (see Procedure for more details). Second, the inclusion of prototypes of sadness and fear—which are typically harder to identify accurately than is anger in the voice—made the task more difficult (Johnstone & Scherer, 2000; Morningstar, Ly, Feldman, & Dirks, 2018; Scherer, 2019). This added complexity provided a rigorous test of whether the shifting effects observed in Experiment 1 would continue to be observed beyond the limited, single stimulus condition in Experiment 1, providing both a replication and extension of those data. Thus, Experiment 2 allowed us to test whether participants continue to track the distributions of vocal cues under more complex conditions involving multiple speakers and vocalization categories.

Method

Participants

Forty adults (32 female; age range = 18–21 years, Mage = 18.84 years, SDage = 0.74 years) participated in this experiment. We aimed for 40 participants based on sample sizes of previous research that also examined these effects (Experiments 2 and 3 in Plate et al., 2019). As we found no age-related differences in how adults and children used distributional information in Experiment 1, we tested only adults in Experiment 2. All participants were undergraduate students at the University of Wisconsin-Madison and received course credit (2.50% African American, 30% Asian American, 10% Hispanic, 57.5% White). The Institutional Review Board approved the research.

Stimuli

Stimuli were created identically to those presented in Experiment 1, but included actor portrayals of neutral morphed into categories of anger, sadness, and fear. Original recordings were produced by one male actor (anger) and two female actors (sadness, fear) saying the sentence "Why did you do that?" in both a neutral and an emotional tone of voice. Morphs were created using the same procedure outlined above, yielding 11 recordings ranging from 10% emotional intensity to 100% emotional intensity for each speaker/emotion. Validation data from 190 listeners suggests listeners were increasingly able to identify the intended emotion in morphs as the emotional intensity increased (Morningstar et al., under review), with accuracy ranging from 18% for the 10% intensity recordings to 59% for the 100% intensity recordings (in a task where chance was 14% accuracy). Nonverbal stimuli were not included in Experiment 2 as we were unable to create a realistic sounding morph for sadness.

Procedure

The procedure was identical to Experiment 1 except Shift Type became a within-participant manipulation. Participants were introduced to three different actors ("John," "Jane," and "Anna"—Actors 24, 10, and 6 from the MacArthur Network Face Stimuli Set; Tottenham et al., 2009), and the number of trials in each phase was adjusted. The task included 72 trials during the training phase (24 per actor) and 144 trials during the testing phase (48 per actor); during testing participants

were exposed to all three emotions and shift conditions (calm shifted, unshifted, or upset shifted). Each shift condition was applied to one of the emotions/actors: for instance, a participant could hear calm shifted expressions of sadness by "Anna," unshifted expressions of anger by "John," and upset shifted expressions of fear by "Jane." The way in which each emotion was shifted was counterbalanced across participants. As in Experiment 1, participants categorized each morph as upset (by clicking the red room) or calm (by clicking the blue room). We kept the rooms and instructions the same as in Experiment 1, as all three emotions are negatively valenced and fall into the "upset" category (even though they may differ in other ways such as arousal levels).

Results: Do Perceivers Shift Their Categorization of Verbal Vocalizations Based upon Different Distribution of Stimuli They Encounter for Multiple Emotions?

Training Phase We analyzed whether participants continued to learn the category boundary during training when tracking three different actors displaying three different emotions. Participants were successful, with an average accuracy of 85.97% during training. We then regressed participants responses (0 = "calm," 1 = "upset") on Percent Upset using a logistical generalized mixed-effects model and again found that for each 10% increase in emotion intensity, participants are more likely to categorize an auditory cue as upset across all emotion types, $b = 1.12$, $\chi^2(1) = 235.46$, $p < .001$, OR = 3.06.

Testing Phase Next, we analyzed whether participants updated their category boundaries for the three different emotions based upon exposure to different distributions. Recall that in Experiment 1 we manipulated shift condition between-subject, while in Experiment 2 we manipulated shift condition within-subject. Still, as in Experiment 1 and in support of our hypothesis, we found that participants shifted their category boundary for each voice identity based on the distributions encountered, $\chi^2(2) = 25.92$, $p < .001$ (see Fig. 3). Calm shifted emotions were identified as "upset" marginally earlier in the morph continuum, $b = 0.83$, $z = 1.76$, $p = .078$, OR = 2.29, and upset shifted emotions were identified as "upset" later in the morph continuum, $b = -1.29$, $z = -3.47$, $p < .001$, OR = 0.27. Post hoc analyses for each of the different emotions are presented in the Supplemental Material.

Discussion of Experiment 2

We replicated and extended the findings of Experiment 1: Participants shifted their categorization of different auditory cues for multiple speakers and emotions after exposure to different distributions. Crucially, participants were able to track this information for multiple categories (and individuals) at once. These findings suggest that perceivers are able to account for individual differences in expressivity in their judgments, and that the general patterns of learning that we observed in Experiment 1 also emerged with a new set of stimuli and emotion cues. However, the current data cannot determine how much individuals were adjusting to different speakers versus different emotions. Overall, Experiment 2 is consistent with the view that these shifts generally impact perceptual learning of vocalizations of emotions.

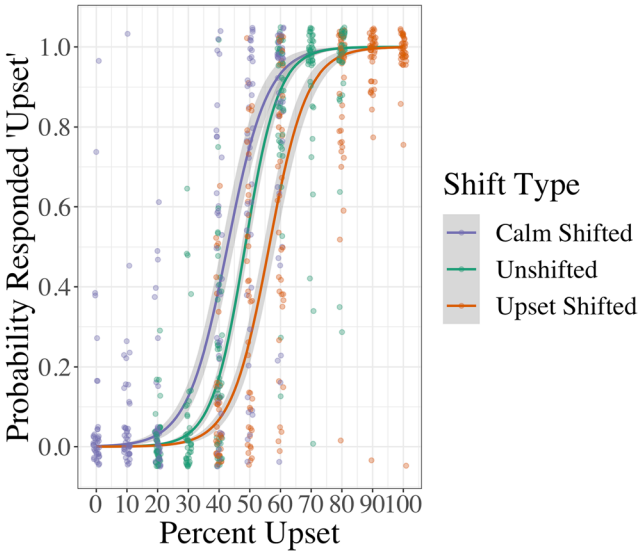

Fig. 3 Multiple emotion testing phase: exposure to varying distributions of verbal auditory stimuli for multiple emotions affected participants' categorization

General Discussion

The present experiments examined whether individuals utilized the distributional properties of perceptual stimuli to flexibly adjust vocal emotion categories. We tasked adults and children with categorizing the affective states communicated by verbal and nonverbal vocalizations that continuously varied from "calm" to "upset" as we varied the distribution of the intensity of the stimuli they encountered. We found participants rapidly adjusted their categorization of auditory emotion cues based upon the statistical distribution of the input to which they were exposed. When a speaker's vocal intensity is limited such that they never express "maximal" negative arousal (as in the calm shifted condition), vocalizations that were previously categorized as calm are categorized as upset. In other words, when listening to less expressive speakers, people have lower thresholds for detecting emotion in the voice. Likewise, if a speaker's expressive range is more intense (as in the upset shifted condition), vocalizations that were previously categorized as upset are categorized as calm. Listeners adapt to highly expressive speakers by increasing

their threshold for detecting emotion in the voice. This adjustment occurred across a range of vocal stimuli (both verbal and nonverbal) and for multiple speakers and negatively valenced emotions. Categorization of auditory cues of emotion appears to be flexible and sensitive to the expressivity of the speaker, as participants rapidly adjust their categorization processes.

In combination with prior research on the categorization of facial cues meant to represent anger (Plate et al., 2019), these results provide evidence for a general learning mechanism that allows children and adults' to adjust to the ways that different people communicate their emotions. This mechanism may be what allows individuals to learn to appropriately respond to social cues despite individual differences and cultural variation in overall expressivity (Laukka & Elfenbein, 2021; Rychlowska et al., 2015; Wood et al., 2016), and even play a role in helping children to learn emotion categories. However, the short-term manipulations of vocal expressivity in the present experiments are not expected to have long-term effects on people's category knowledge. Future research could investigate whether repeated exposure to different distributions, for instance being socialized in families or cultures with different expressive norms, creates stable individual differences in how vocalizations are interpreted—and how these distributions interact with more instantaneous summary statistics (Whitney & Yamanashi Leib, 2018). Such data could also reveal how differences in intensity might influence ratings of speaker characteristics, or how a participant's adjustment to speaker expressivity contributes to empathic accuracy (Zaki et al., 2008). Here, we examined categorical ratings, but there is some suggestion in recent research that continuous ratings are also likely changed through exposure to different distributions of information (Leitzke et al., 2020).

The similarities between our findings and other domains, such as speech perception (Samuel & Kraljic, 2009; Weatherholtz & Jaeger, 2016), presents an opportunity to utilize models and research in these areas. For instance, models of speech perception suggest that speakers track and use variation across speaker groups if that information is informative and useful. In speech perception, variables like age, gender or dialect may aid speech categorization in situations where these variables reliably predict patterns of speech variability (Kleinschmidt, 2019; Kleinschmidt & Jaeger, 2015; Kleinschmidt & Jaeger, 2011). It is not feasible (and likely unhelpful) for perceivers to track all possible sources of variability in how different individuals convey emotion, but it is possible that perceivers use social groupings in ways that are similar to speech models—such as age, gender, and perceived regional/cultural background—as potentially salient cues when tracking variation in emotional expressivity.

Does the variability in how emotions are conveyed diminish the role of such surface features in emotion learning? It is likely that perceptual features related to emotion are so variable, that children may need to rely on language and other converging cues, in addition to facial and vocal information, to learn these categories (Hoemann et al., 2019). However, the role of perceptual features commonly associated with emotion categories is also a critical piece of this learning puzzle (Keltner et al., 2019). In our view, delineating boundaries between conceptual versus perceptual effects in emotion fails to account for the ways in which perceptions and concepts overlap and influence each other (Goldstone & Barsalou, 1998). For instance, labels help to guide infant learning, but only if those labels correlate with perceptual features (Plunkett, 2011; Plunkett et al., 2008). Similarly, there are many examples of the ways in which variability of perceptual input allows children to meaningfully separate tokens as the basis for formulating relevant categories (Adriaans & Swingley, 2017). The present data suggest that theories of emotion need to adequately consider the role of early perceptual input in the formation of emotion concepts, while also accounting for the role that perceptual experience plays in the formation of those concepts and categories. In these ways, it is the very natural variation in how emotions are communicated that may be an important source of how children learn emotion categories.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s42761-021-00038-w.

Additional Information

Acknowledgements We thank Andrey Anikin for creating the nonverbal vocal stimuli using his open-source tool, Soundgen, the families who participated in this study, and the research assistants who helped conduct the research, especially Sarah Fieweger and Quentin Wedderburn.

Funding Information Funding for this project was provided by the National Institute of Mental Health (MH61285) to S. Pollak and a core grant to the Waisman Center from the National Institute of Child Health and Human Development (U54 HD090256). K. Woodard was supported by a University of Wisconsin Distinguished Graduate Fellowship.

Data Availability Stimuli, data, task scripts, and R scripts are available online at https://osf.io/749xq/?view_only=ef7f9d9509284ed2927948509c596db5. Although the participants who produced certain vocal stimuli did not consent to sharing the stimuli publicly online, they did give us permission to share the stimuli with other researchers upon request. The experiments were not preregistered.

Conflict of Interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

Ethics Approval All studies in this manuscript were approved by the Education and Social/Behavioral Science Institutional Review Board at the University of Wisconsin – Madison.

Informed Consent All participants gave informed consent prior to their participation in the study.

References

Adriaans, F., & Swingley, D. (2017). Prosodic exaggeration within infant-directed speech: consequences for vowel learnability. *The Journal of the Acoustical Society of America, 141*(5), 3070–3078.

Aguert, M., Laval, V., Lacroix, A., Gil, S., & Le Bigot, L. (2013). Inferring emotions from speech prosody: not so easy at age five. *PLoS ONE, 8*(12), e83657. https://doi.org/10.1371/journal.pone.0083657.

Anikin, A. (2019). Soundgen: an open-source tool for synthesizing nonverbal vocalizations. *Behavior Research Methods, 51*(2), 778–792. https://doi.org/10.3758/s13428-018-1095-7.

Anikin, A., & Persson, T. (2017). Nonlinguistic vocalizations from online amateur videos for emotion research: a validated corpus. *Behavior Research Methods, 49*(2), 758–771. https://doi.org/10.3758/s13428-016-0736-y.

Banse, R., & Scherer, K. R. (1996). Acoustic profiles in vocal emotion expression. *Journal of Personality and Social Psychology, 70*(3), 614–636. https://doi.org/10.1037/0022-3514.70.3.614.

Barrett, L. F., Adolphs, R., Marsella, S., Martinez, A. M., & Pollak, S. D. (2019). Emotional expressions reconsidered: challenges to inferring emotion from human facial movements. *Psychological science in the public interest, 20*(1), 1–68.

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

Clarke, C. M., & Garrett, M. F. (2004). Rapid adaptation to foreign-accented English Rapid adaptation to foreign-accented English. *The Journal of the Acoustical Society of America, 116*(6), 3647–3658. https://doi.org/10.1121/1.1815131.

Clarke-Davidson, C. M., Luce, P. A., & Sawusch, J. R. (2008). Does perceptual learning in speech reflect changes in phonetic category representation or decision bias? *Perception and Psychophysics, 70*(4), 604–618. https://doi.org/10.3758/PP.70.4.604.

Core Team, R. (2019). R: a language and environment for statistical computing. In *R Foundation for Statistical Computing*. Vienna: Austria. URL https://www.R-project.org/.

Dahan, D., Drucker, S. J., & Scarborough, R. A. (2008). Talker adaptation in speech perception: adjusting the signal or the representations? *Cognition, 108*(3), 710–718. https://doi.org/10.1016/j.cognition.2008.06.003.

Dotsch, R., Hassin, R. R., & Todorov, A. (2017). Statistical learning shapes face evaluation. *Nature Human Behaviour, 1*(1), 1–6. https://doi.org/10.1038/s41562-016-0001.

Friend, M., & Bryant, J. B. (2000). A developmental lexical bias in the interpretation of discrepant messages. *Merrill Palmer Q (Wayne State Univ Press), 46*(2), 342–369 PMID: 28698709; PMCID: PMC5502114.

Frost, R., Armstrong, B. C., & Christiansen, M. H. (2019). Statistical learning research: a critical review and possible new directions. *Psychological Bulletin, 145*(12), 1128–1153. https://doi.org/10.1037/bul0000210.

Frost, R., Armstrong, B. C., Siegelman, N., & Christiansen, M. H. (2015). Domain generality versus modality specificity: the paradox of statistical learning. *Trends in Cognitive Sciences. Elsevier Ltd., 19*, 117–125. https://doi.org/10.1016/j.tics.2014.12.010.

Goldstone, R. L., & Barsalou, L. W. (1998). Reuniting perception and conception. *Cognition, 65*(2-3), 231–262 conception.

Green, P., & MacLeod, C. J. (2016). SIMR: an R package for power analysis of generalized linear mixed models by simulation. *Methods in Ecology and Evolution, 7*(4), 493–498. https://doi.org/10.1111/2041-210X.12504.

Hawk, S. T., Van Kleef, G. A., Fischer, A. H., & Van Der Schalk, J. (2009). "Worth a thousand words": absolute and relative decoding of nonlinguistic affect vocalizations. *Emotion, 9*(3), 293.

Hoemann, K., Xu, F., & Barrett, L. F. (2019). Emotion words, emotion concepts, and emotional development in children: a constructionist hypothesis. *Developmental psychology, 55*(9), 1830.

Ito, K. (2018). Gradual development of focus prosody and affect prosody comprehension. *The Development of Prosody in First Language Acquisition, 23*, 247.

Johnson, K. (2005). Speaker normalization in speech perception. The handbook of speech perception. In D. B. Pisoni & R. Remez (Eds.), *The handbook of speech perception* (pp. 363–389). Oxford: Blackwell Publishers https://doi.org/9780470757024.

Johnstone, T., & Scherer, K. (2000). *Vocal communication of emotion. Handbook of emotions* (Vol. *2*, pp. 220–235) Retrieved from https://www.researchgate.net/publication/248349015.

Juslin, P. N., & Laukka, P. (2001). Impact of intended emotion intensity on cue utilization and decoding accuracy in vocal expression of emotion. *Emotion, 1*(4), 381–412. https://doi.org/10.1037/1528-3542.1.4.381.

Kawahara, H., & Morise, M. (2011). Technical foundations of TANDEM-STRAIGHT, a speech analysis, modification and synthesis framework. *Sadhana, 36*(5), 713–727.

Kawahara, H., Morise, M., Takahashi, T., Nisimura, R., Irino, T., & Banno, H. (2008). Tandem-STRAIGHT: a temporally stable power spectral representation for periodic signals and applications to interference-free spectrum, F0, and aperiodicity estimation. In *2008 IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 3933–3936) IEEE.

Keltner, D., Tracy, J. L., Sauter, D., & Cowen, A. (2019). What basic emotion theory really says for the twenty-first century study of emotion. *Journal of nonverbal behavior, 43*(2), 195–201.

Kleinschmidt, D., & Jaeger, T. F. (2011). *A Bayesian belief updating model of phonetic recalibration and selective adaptation*. Association for Computational Linguistics.

Kleinschmidt, D. F. (2019). Structure in talker variability: how much is there and how much can it help? *Language. Cognition and Neuroscience, 34*(1), 43–68. https://doi.org/10.1080/23273798.2018.1500698.

Kleinschmidt, D. F., & Jaeger, T. F. (2015). Robust speech perception: recognize the familiar, generalize to the similar, and adapt to the novel. *Psychological Review, 122*(2), 148–203. https://doi.org/10.1037/a0038695.

Laukka, P., & Elfenbein, H. A. (2021). Cross-cultural emotion recognition and in-group advantage in vocal expression: a meta-analysis. *Emotion Review, 1754073919897295*.

Leitzke, B. T., Plate, R. C., & Pollak, S. D. (2020). Training reduces error in rating the intensity of emotions. *Emotion*.

Levari, D. E., Gilbert, D. T., Wilson, T. D., Sievers, B., Amodio, D. M., & Wheatley, T. (2018). Prevalence-induced concept change in human judgment. *Science, 360*(6396), 1465–1467. https://doi.org/10.1126/science.aap8731.

Liberman, A. M. (1957). Some results of research on speech perception. *The Journal of the Acoustical Society of America, 29*(1), 117–123.

Liberman, A. M., Cooper, F. S., Shankweiler, D. P., & Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological Review, 74*(6), 431–461. https://doi.org/10.1037/h0020279.

Lüdecke D (2020). *sjPlot: data visualization for statistics in social science*. R package version 2.8.6, https://CRAN.R-project.org/package=sjPlot.

Maye, J., Werker, J. F., & Gerken, L. (2002). Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition, 82*(3), B101–B111.

Miller, J. L., & Eimas, P. D. (1995). *Speech perception: from signal to word*. *Annual Review of Psychology, 46*(1), 467–492 Retrieved from www.annualreviews.org.

Morningstar, M., Dirks, M. A., & Huang, S. (2017). Vocal cues underlying youth and adult portrayals of socio-emotional expressions. *Journal of Nonverbal Behavior, 41*(2), 155–183.

Morningstar, M., Ly, V. Y., Feldman, L., & Dirks, M. A. (2018). Mid-adolescents' and adults' recognition of vocal cues of emotion and social intent: differences by expression and speaker age. *Journal of Nonverbal Behavior, 42*(2), 237–251. https://doi.org/10.1007/s10919-018-0274-7.

Morningstar, M., Nelson, E. E., & Dirks, M. A. (2018). Maturation of vocal emotion recognition: insights from the developmental and neuroimaging literature. *Neuroscience and Biobehavioral Reviews. Elsevier Ltd., 90*, 221–230. https://doi.org/10.1016/j.neubiorev.2018.04.019.

Morton, J. B., & Trehub, S. E. (2001). Children's understanding of emotion in speech. *Child development, 72*(3), 834–843.

Newman, R. S., & Sawusch, J. R. (1996). Perceptual normalization for speaking rate: effects of temporal distance. *Perception and Psychophysics, 58*(4), 540–560. https://doi.org/10.3758/bf03213089.

Plate, R. C., Wood, A., Woodard, K., & Pollak, S. D. (2019). Probabilistic learning of emotion categories. *Journal of Experimental Psychology: General, 148*(10), 1814–1827. https://doi.org/10.1037/xge0000529.

Plate, R. C., Woodard, K., & Pollak, S. D. (in press). Statistical learning in an emotional world. In D. Dukes, A. C. Samson, & E. A. Walle (Eds.), *The Oxford handbook of emotional development*. Oxford: Oxford University Press.

Plunkett, K. (2011). The role of auditory stimuli in infant categorization. *Infant perception and cognition: recent advances, emerging theories, and future directions*, 203–221.

Plunkett, K., Hu, J. F., & Cohen, L. B. (2008). Labels can override perceptual categories in early infancy. *Cognition, 106*(2), 665–681.

Rychlowska, M., Miyamoto, Y., Matsumoto, D., Hess, U., Gilboa-Schechtman, E., Kamble, S., Muluk, H., Masuda, T., & Niedenthal, P. M. (2015). Heterogeneity of long-history migration explains cultural differences in reports of emotional expressivity and the functions of smiles. *Proceedings of the National Academy of Sciences of the United States of America, 112*(19), E2429–E2436. https://doi.org/10.1073/pnas.1413661112.

Saffran, J. R. (2020). Statistical language learning in infancy. *Child Development Perspectives, 14*(1), 49–54. https://doi.org/10.1111/cdep.12355.

Samuel, A. G., & Kraljic, T. (2009). Perceptual learning for speech. *Attention, Perception, and Psychophysics, 71*(6), *1207–1218*. https://doi.org/10.3758/APP.71.6.1207.

Sauter, D. A., Eisner, F., Calder, A. J., & Scott, S. K. (2010). Perceptual cues in nonverbal vocal expressions of emotion. *Quarterly Journal of Experimental Psychology, 63*(11), 2251–2272. https://doi.org/10.1080/17470211003721642.

Sauter, D. A., Panattoni, C., & Happé, F. (2013). *Children's recognition of emotions from vocal cues* (pp. 97–113). https://doi.org/10.1111/j.2044-835X.2012.02081.x.

Scarantino, A. (2014). Basic emotions, psychological construction and the problem of variability. In L. F. Barrett & J. A. Russell (Eds.), *The psychological construction of emotion* (pp. 334–376). New York: Guilford Press.

Scherer, K. R. (2019). Acoustic patterning of emotion vocalizations. In Frühholz & P. Belin (Eds.), *The Oxford handbook of voice perception* (pp. 61–92). Oxford: Oxford University Press.

Siegelman, N., Bogaerts, L., Elazar, A., Arciuli, J., & Frost, R. (2018). Linguistic entrenchment: prior knowledge impacts statistical learning performance. *Cognition, 177*, 198–213. https://doi.org/10.1016/j.cognition.2018.04.011.

Siegelman, N., & Frost, R. (2015). Statistical learning as an individual ability: theoretical perspectives and empirical evidence. *Journal of memory and language, 81*, 105–120.

Tottenham, N., Tanaka, J. W., Leon, A. C., McCarry, T., Nurse, M., Hare, T. A., Marcus, D. J., Westerlund, A., Casey, B. J., & Nelson, C. (2009). The NimStim set of facial expressions: judgments from untrained research participants. *Psychiatry Research, 168*(3), 242–249. https://doi.org/10.1016/j.psychres.2008.05.006.

Weatherholtz, K. (2015). *Perceptual learing of systemic cross-category vowel variation*. *Doctoral Dissertation*: *The Ohio State University*.

Weatherholtz, K., & Jaeger, T. F. (2016). *Speech perception and generalization across talkers and accents*. https://doi.org/10.1093/ACREFORE/9780199384655.013.95.

Whitney, D., & Yamanashi Leib, A. (2018). Ensemble perception. *Annual review of psychology, 69*, 105–129.

Wickham, H. (2016). *ggplot2: elegant graphics for data analysis*. Berlin: Springer.

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D. A., François, R., et al. (2019). Welcome to the Tidyverse. *Journal of Open Source Software, 4*(43), 1686.

Wood, A., Rychlowska, M., & Niedenthal, P. M. (2016). Heterogeneity of long-history migration predicts emotion recognition accuracy. *Emotion, 16*(4), 413–420. https://doi.org/10.1037/emo0000137.

Xie, X., Weatherholtz, K., Bainton, L., Rowe, E., Burchill, Z., Liu, L., & Jaeger, T. F. (2018). Rapid adaptation to foreign-accented speech and its transfer to an unfamiliar talker. *The Journal of the Acoustical Society of America, 143*(4), 2013–2031. https://doi.org/10.1121/1.5027410.

Zaki, J., Bolger, N., & Ochsner, K. (2008). It takes two: the interpersonal nature of empathic accuracy. *Psychological Science, 19*(4), 399–404.

Zhang, Y., Hedo, R., Rivera, A., Rull, R., Richardson, S., & Tu, X. M. (2019). Post hoc power analysis: is it an informative and meaningful analysis? *General Psychiatry, 32*(4), e100069.