Beyond Words: Predicting Preference Strength From Vocal Cues

Maximilian Gaerth
Hauke Roggenkamp
Christian Hildebrand
Zhenling Jiang

Maximilian Gaerth (corresponding author) is Assistant Professor of Marketing, Goizueta Business School, Emory University, USA (email: mgaerth@emory.edu). Hauke Roggenkamp is Research Associate and PhD Candidate, Institute of Behavioral Science and Technology, University of St. Gallen, Switzerland (email: hauke.roggenkamp@unisg.ch). Christian Hildebrand is Full Professor of Marketing Analytics and Director, Institute of Behavioral Science and Technology, University of St. Gallen, Switzerland (email: christian.hildebrand@unisg.ch). Zhenling Jiang is Assistant Professor of Marketing, The Wharton School, University of Pennsylvania, USA (email: zhenling@wharton.upenn.edu). Please address correspondence to Maximilian Gaerth.
The authors thank the Wharton Behavioral Lab for funding and supporting this research.

Beyond Words: Predicting Preference Strength From Vocal Cues

Abstract

Consumers increasingly make purchase decisions through voice-enabled technologies, from shopping via smart speakers at home to AI-powered ordering kiosks at fast-food restaurants. In these interactions, consumers not only communicate their choices through *what* they say but also through *how* they say it. The current work reveals that so-called vocal paralinguistic cues, such as a person's vocal tone and intonation, provide an unexplored window into consumers' underlying states at the moment of a decision. Using gradient-boosted decision trees, we introduce Voice-Inferred Preference Strength (VIPS) and demonstrate that vocal paralinguistic cues reliably predict consumers' preference strength and behaviorally relevant outcomes. We further demonstrate that the approach generalizes across a wide array of choice settings (from product preferences to beauty ratings). This research provides firms with a novel lens into consumers' preferences and decision states at the point of purchase. We also make our model publicly available for interested researchers and provide a dedicated Shiny app to use the model as a feature-extraction tool. We conclude by highlighting the theoretical and practical implications of our findings in an increasingly voice-first marketplace.

Keywords: voice, voice analytics, voice commerce, vocal paralanguage, preference strength, machine learning

Most of the data generated in today's economy is unstructured. By some estimates, between 80% and 95% of all business data lacks a predefined structure, with the vast majority representing text data (Gandomi and Haider 2015). Over the past two decades, marketing scholars have demonstrated the remarkable value of such unstructured text data for understanding and predicting consumer behavior. For instance, written language used in marketplace interactions has been shown to reflect both stable traits (e.g., personality, values; Mogilner, Aaker, and Kamvar 2011; Moon and Kamakura 2017) and transient states (e.g., moods, attitudes; Ludwig et al. 2016; Rocklage et al. 2023; Rude, Gortner, and Pennebaker 2004), and to predict behaviorally relevant outcomes such as loan default, deception, and product adoption (e.g., Anderson and Simester 2014; Hancock et al. 2007; Netzer, Lemaire, and Herzenstein 2019; Ott, Cardie, and Hancock 2012). Taken together, this prior work has established text as a powerful window into consumers' psychology and intentions.

Yet consumer-firm interactions are no longer dominated by text traces alone. Advances in artificial intelligence and speech recognition have made voice-enabled interactions increasingly central to commerce. Voice commerce was valued at US$ 45.75 billion in 2023, and is projected to reach US$ 186.28 billion by 2030, growing at a 24.6% compound annual growth rate (Grand View Research 2024). Smart speakers alone (e.g., Amazon Echo, Google Nest, etc.) account for nearly half of all voice commerce transactions (market.us 2025), while major chains such as Taco Bell and McDonald's process millions of orders via fully automated voice interfaces (Darley 2025; Mira 2024).

This raises an important methodological and substantive question: beyond *what* consumers say, can also *how* they say it also reveal meaningful, marketing-relevant information about their underlying decision states? And can the sound of people's voices when expressing their preferences be used to predict behaviorally relevant outcomes? Specifically, we focus on

vocal paralanguage, which refers to the non-verbal elements of speech, including basic voice characteristics of a speaker (e.g., the frequency or pitch, loudness, or timbre), speech characteristics (e.g., tempo and pauses during speech formation), and nonspeech sounds (e.g., breathing sounds or laughter). Unlike text data, vocal paralanguage is largely involuntary, difficult to revise (Cummins, Baird, and Schuller 2018; Scherer, Johnstone, and Klasmeyer 2003), and a reliable window into consumers' cognitive and affective states (Jurafsky and Martin 2021; Scherer 1986; Scherer, Johnstone, and Klasmeyer 2003).

The current research examines one particularly relevant decision state in which consumers' cognitive and affective experience matters: their preference strength when making a decision. Consumers' preference strength generally describes the confidence with which consumers choose one option over other options (Konovalov and Krajbich 2019). Understanding consumers' preference strength matters as a critical antecedent to choice deferral (Dhar 1997), willingness to pay (Spiller and Belogolova 2017), and loyalty (Yoon and Simonson 2008). However, existing approaches for measuring preference strength face inherent limitations. Self-reports typically suffer from social desirability biases and people's lack of introspective access to their underlying cognitive or affective states, processes, or motivations (i.e., introspection; Nisbett and Wilson 1977; Podsakoff et al. 2003). Moreover, expressing oneself through written responses can prompt respondents to come up with post hoc rationalizations rather than reflecting real-time decision processes (Pham 2013; Pham et al. 2015). The current research introduces Voice-Inferred Preference Strength (VIPS) and demonstrates that directly modeling consumers' vocal paralanguage offers a naturally occurring, real-time, and unobtrusive measure of consumers' preference strength where it matters most—during consumers' actual product choices and purchase decisions.

We test this proposition using a machine learning approach applied to vocal paralinguistic features extracted from consumer speech. Across four studies, we demonstrate that vocal paralanguage reliably predicts consumers' preference strength and managerially relevant outcomes. These findings make four primary contributions: First, we establish vocal paralanguage as a novel approach to measure consumer preference strength beyond traditional measurement approaches (e.g., self-reports using Likert-scales, open-text data, or mere response times; see Rocklage et al. 2023). Second, the current work provides a nuanced lens to extend our understanding of how decisional conflict manifests in consumers' speech data. We show that the cognitive and affective processes underlying consumer preference formation systematically "leak" into consumers' vocal expressions in detectable ways. Third, we demonstrate that voice-derived measures of preference strength predict behaviorally relevant outcomes, including consumers' willingness to pay a price premium. Fourth, and importantly, our custom web interface built in oTree (Chen, Schonger, and Wickens 2016) represents a practical contribution in its own right: it enables scalable, high-quality collection of labeled voice data in controlled settings. This infrastructure, along with our unique dataset, offers a benchmark for developing and testing new methods, supporting replication, and fostering innovation at the intersection of marketing, decision-making, and computational linguistics.

The remainder of this paper is organized as follows. We first review the theoretical foundations linking vocal paralanguage to consumers' preference strength. We then provide an overview of our data generation and modeling process before presenting four empirical studies designed to develop, validate, and assess the generalizability of our approach across choice contexts. We conclude by discussing theoretical and practical implications for an increasingly voice-first marketplace and providing directions for future work.

Theoretical Background

Vocal Paralanguage as Involuntary Consumer Data

When consumers make a choice, they engage in a dynamic process of resolving decisional conflict between competing options (Bettman and Sujan 1987; Malkoc, Zauberman, and Ulu 2005). Strong preferences entail little conflict because one option clearly dominates, whereas weak preferences involve sustained conflict as consumers deliberate between similarly attractive alternatives (see drift-diffusion models; e.g., Ratcliff et al. 2016). We argue that this internal conflict systematically leaks into consumers' vocal paralanguage—the nonverbal vocal features that accompany human speech, including noticeable pitch variations, tempo changes, such as audible hesitations, and voice quality fluctuations (Schuller and Batliner 2013). Unlike the words consumers say, which are typically under conscious control, vocal paralanguage emerges largely involuntarily from the physiological and psychological states underlying people's decision-making (Cummins, Baird, and Schuller 2018; Melzner, Bonezzi, and Meyvis 2023; Scherer, Johnstone, and Klasmeyer 2003).

Just as motor movements have been shown to reveal these dynamics in real time (Song and Nakayama 2009; Spivey 2008), vocal paralanguage should provide a continuous, real-time trace of how conflicted consumers are when making a choice. Another reason these vocal paralinguistic cues are especially diagnostic for inferring a consumer's preference strength is that speech is inherently temporal and irreversible. Vocalizations emerge spontaneously in real time, without the opportunity to revise it (Zierau et al. 2022). As such, vocal paralanguage is more likely to reflect the evolving dynamics of mental states during decision-making than the post hoc rationalizations typically found in written self-reports. In line with previous work suggesting that stronger preferences correspond with lower conflict (Konovalov and Krajbich 2019; Stillman and

Ferguson 2019), we expect that the evaluative process when making a choice between alternatives should systematically influence people's speech production. Specifically, strong preferences should be reflected by more confident and decisive vocal patterns, whereas weak preferences should be reflected by more pronounced vocal hesitation and less stable pitch and tempo (i.e., irregular speaking styles and more inconsistent pacing that reflect the typical hesitation accompanied by decision conflict)—consistent with conceptual and qualitative observations from the persuasion and sales literature (e.g., Van Zant and Berger 2019).

Vocal Paralanguage in Comparison to Existing Preference Strength Measures

Current approaches to measuring consumers' preference strength capture primarily voluntary and explicit aspects of consumers' underlying cognitive and affective experience but only partial information about their idiosyncratic internal states at the moment of decision. Table 1 provides a summary and comparison of the predominant techniques used in marketing and related fields. Historically, self-reports via Likert-scales have been the dominant measurement technique for preference strength due to their ease of use and administration (e.g., Chernev 2001; Fisher and Woolley 2024; Khan and Dhar 2006; Nowlis and Simonson 1997; Song, Jung, and Zhang 2021). However, a major limitation is that such measures are subject to social desirability biases and consumers' limited ability to fully articulate their internal states (Kleiman and Hassin 2011).

On the other hand, prior work successfully developed text-based measures to offer more implicit insights (e.g., linguistic patterns used to infer mental states such as uncertainty; Rocklage et al. 2023). However, written responses are often consciously composed and edited, reflecting the outcome of deliberation rather than the decision process as it unfolds.

Another widely used approach is the analysis of response time (RT)—the latency between the presentation of a stimulus (e.g., a choice option, question, or task) and the individual's

behavioral response (e.g., a decision, keypress, or verbal answer). RT has been widely adopted as a relatively unobtrusive proxy for preference strength (Konovalov and Krajbich 2019). Although RT is a continuous measure and may therefore capture more nuanced information than discrete choice outcomes (Loomes 2005; Spiliopoulos and Ortmann 2018), it is often contaminated by "nondecision time," such as reading, perceptual encoding, or motor latency (Luce 1986). This makes RT a noisy proxy for preference strength in marketplace contexts. Moreover, RT remains fundamentally one-dimensional and limited to a single data point at the end of the deliberation. As such, RT cannot reveal the dynamic cognitive and affective processes unfolding during the decision itself.

To address these limitations, researchers have recently turned to mouse-tracking to assess decisional conflict indirectly by analyzing the directness (or lack thereof) of mouse movements toward a target (i.e., the extent to which the unchosen option attracts participants' mouse movements; Freeman 2018; Stillman, Shen, and Ferguson 2018). While mouse-tracking reveals how decisional conflict unfolds in real time, it is constrained by its reliance on deliberate motor actions in laboratory settings and is impractical in many consumer–firm interactions (e.g., retail), which are increasingly touch- and voice-driven.

Taken together, existing approaches to inferring preference strength either rely on conscious post hoc reports or capture a narrow slice of the decision process. What is missing is a natural and real-time measure of preference strength that can operate in the contexts where consumers actually interact with firms. We propose that vocal paralanguage offers precisely this opportunity. Paralinguistic features can be captured during consumer-firm interactions without requiring special equipment or explicit awareness from consumers. We therefore test whether consumers' vocal paralinguistic features can reliably predict preference strength and whether

these predictions correlate with behaviorally relevant outcomes, such as willingness to pay a price premium.

Finally, it is worth highlighting that directly modeling consumers' vocal paralanguage also offers important practical advantages compared to established physiological measures such as pupil dilation or skin conductance (e.g., Cittadini et al. 2023). Whereas such physiological measures typically require specialized equipment and tightly controlled settings, vocal paralinguistic cues can be captured unobtrusively during natural marketplace interactions. This enables researchers and firms to detect implicit signals of preference strength without interrupting the decision-making process or making participants aware of the measurement apparatus—thereby also reducing potential demand effects.

Table 1: Characteristics of Preference Strength Measurement Methods

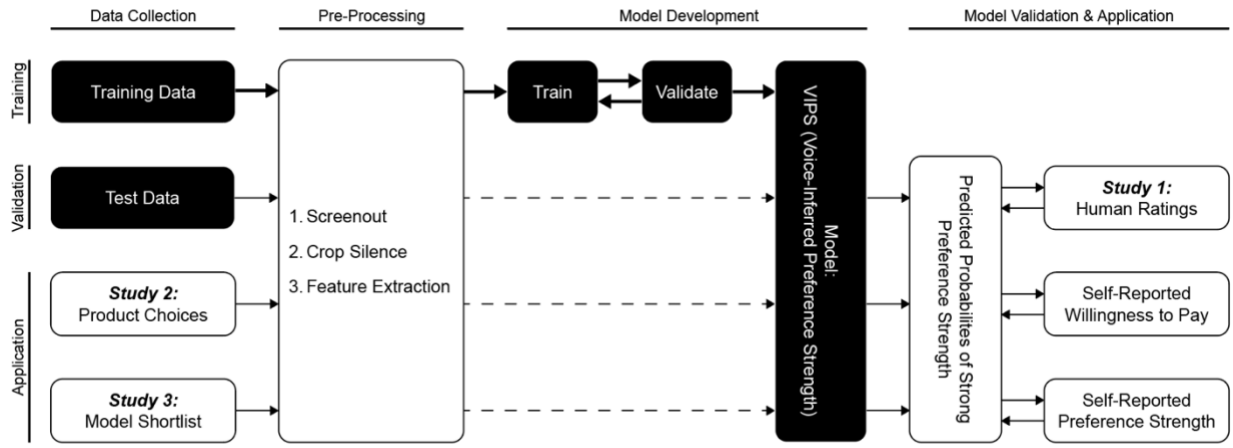
	Voice	Text	Response Time	Mouse Tracking
Involuntariness	Moderate	Low	High	High
Revisability	Low	High	Low	Low
Spontaneity	High	Moderate	High	Moderate
Depth	Moderate	High	Low	Low
Dynamism	High	Low	Low	Moderate
Applicability	Broad	Broad	Moderate	Narrow
Obtrusiveness	Low	Moderate	Low	Low
Temporality	High	Moderate	Low	High

Overview of Studies

We present four studies that investigate how vocal paralanguage can be used to predict consumer preference strength. First, using voice actors, we develop a labeled dataset that tightly controls for semantic content (i.e., *what* consumers say) to isolate the effect of vocal paralinguistic cues (i.e., *how* they say it). We then employ a series of machine learning techniques to predict consumers' preference strength from these cues (Study 1) and validate our

computational approach with human raters (Study 2). Finally, we examine downstream applications by testing whether voice-predicted preference strength relates to consumers' willingness to pay across products and brands (Study 3) and their self-reported preference strength (Study 4). Taken together, these studies provide empirical evidence regarding the unique value of paralinguistic cues for predicting consumers' preference strength and the relevance of these cues for marketing research. All studies were conducted using custom-built web interfaces (randomization procedures using oTree [Chen, Schonger, and Wickens 2016] plus additional custom scripts for microphone access and real-time audio capture). All processed data, analysis scripts, and reproduction materials are available in our Open Science Framework repository to ensure computational reproducibility of our findings (see https://osf.io/7k98n/?view_only=aa0fe416ff7941d3bc4d201abfc6e599). Figure 1 provides a summary of the data processing pipeline and analytical workflow across all four studies.

Figure 1: Overview of Model Development, Validation, and Application Across Four Studies



Study 1: Model Development and Training

In this section, we describe how we developed and validated a labelled data set and trained our machine learning model, a gradient boosting classifier (XGBoost), to predict preference strength.

Data Collection

To create a dataset with labeled preference expressions, we recruited 1,112 US-American English native speakers via Prolific to serve as voice actors in a 2 (preference statements) × 3 (preference expression styles: neutral vs. weak vs. strong) within-subjects design. Participants recorded two standardized decision statements: *"I'll go with the product on the right" and "Out of the two services, I am choosing the second one."* These statements were selected because they represented realistic consumer choice scenarios while varying grammatical structure and linguistic complexity.

Each participant was instructed to deliver the two statements in three preference strength styles: natural, strong, and weak. Recordings always began with the natural (baseline) version, while the order of weak and strong preference strength expressions was randomized. To ensure that acoustic variation reflects paralinguistic expression rather than differences in verbal content, participants were required to use the same wording across all recordings. They were also instructed to record in relatively quiet, private environments using their personal devices to minimize background noise (see Web Appendix A for survey instrument).

Data Processing

Screening and Preprocessing

Audio recordings underwent screening and preprocessing before vocal paralinguistic feature extraction. First, we conducted a manual screen out of all participants who did not comply

with study protocols for at least one statement.[1] Next, we preprocessed the remaining audio files using the tuneR package (Ligges et al. 2013) to remove sections of silence. Excluding these non-informative silent periods (1) ensured that analyses focused solely on vocal paralinguistic cues rather than on pauses or background noise surrounding the preference expression, and (2) prevented silence frames surrounding the preference expression from diluting meaningful differences in vocal paralinguistic features, thereby improving model sensitivity. These steps led to a final dataset containing 3,683 recordings (representing both weak and strong preference expressions) from 921 participants (M_{Age} = 43.7, SD_{Age} = 13.8; 49.8% female).

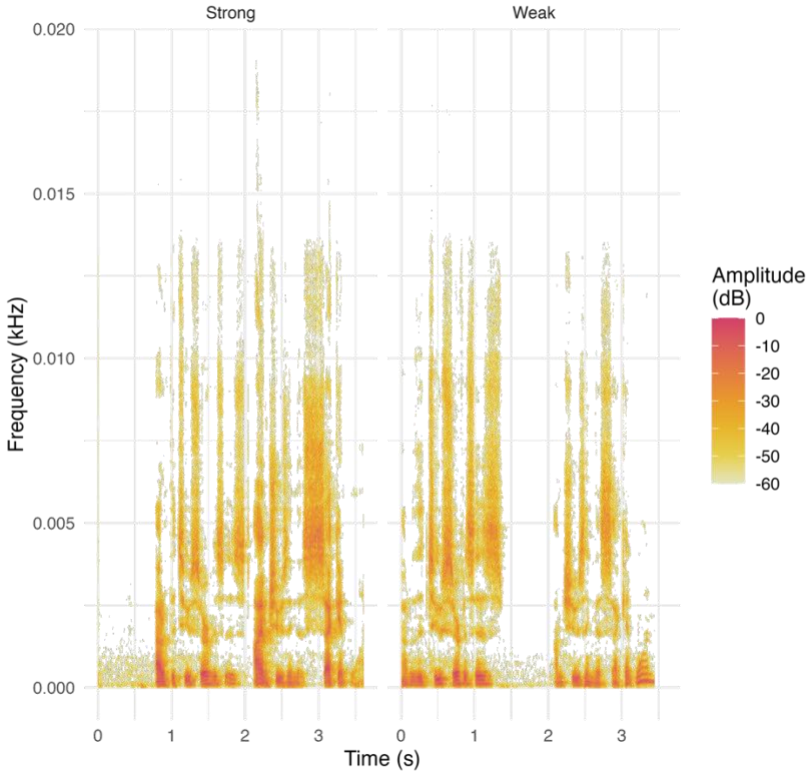
Vocal Paralinguistic Feature Extraction

To extract vocal paralinguistic features from the voice actor recordings, we employed the voiceR package in R (Busquet and Hildebrand 2023), which preprocesses audio files and computes a wide range of acoustic measures. Specifically, we extracted 108 vocal paralinguistic features across four domains: fundamental frequency statistics (e.g., pitch), spectral characteristics (e.g., spectral flux), voice quality measures (e.g., jitter), and temporal features (e.g., speech rate). The audio signal was first transformed into the frequency domain using short-time Fourier transformations, enabling the decomposition of the waveform into frequency-based segments suitable for analysis. Framing segmented the signal into short overlapping windows, which were then recombined to preserve temporal continuity, while windowing reduced spectral leakage due to truncated waveforms. Pre-emphasis filtering was applied to amplify higher-frequency components, thereby enhancing the overall signal-to-noise ratio and minimizing recording-related distortions.

[1] Common violations included deviating from the standardized decision statements, recording in noisy environments, or producing inaudible recordings. This quality control step removed non-compliant participants entirely (across all their recordings)

Figure 2 presents two exemplary spectrograms to illustrate and compare the "vocal footprint" of strong versus weak preference strength expressions for the same participant. A spectrogram displays how the frequency content (i.e., resonance of the voice) and amplitude (i.e., loudness or energy) of speech changes over time. When this person expressed a weak preference compared to a strong preference, the speech signal appears flatter and more monotone (i.e., with less variation in loudness), and shows reduced spectral variability (i.e., fewer distinct frequency bands changing dynamically). In contrast, strong preference speech displays greater spectral variability (i.e., richer fluctuations across frequency bands), clearer harmonic structure (i.e., more consistent and periodic sound waves), and stronger vocal emphasis (i.e., higher peaks of loudness over time). This pattern already suggests visually that weak preference speech patterns appear less dynamic and less harmonic, whereas strong preference speech patterns are more coherent, energetic, and expressive.

Figure 2: Spectrogram of Strong Versus Weak Preferences



Note: Exemplary spectrograms illustrating acoustic variation contrasting weak versus strong preferences. Both panels show the same participant producing the statement "*Out of the two services, I am choosing the second one."* under the two instructed preference strength conditions. The strong preference condition (left) shows bands extending across a wider frequency range and show more irregularity in intensity. In contrast, the weak preference condition (right) shows a more concentrated energy, with less spread into higher frequencies and fewer fluctuations in amplitude across time.

Model Training

We trained a gradient boosting classifier (XGBoost) to predict the instructed preference-strength labels from the extracted paralinguistic voice features. Importantly, since our data includes multiple recordings per participant, we employed grouped five-fold cross-validation using participant identity as the grouping variable. This procedure ensured that recordings from the same participant did not appear in both training and validation sets, thereby preventing data leakage and producing more reliable estimates of model performance.

We tuned the hyperparameters in the model using an exhaustive grid search across 27 parameter combinations, varying the learning rate (0.01, 0.05, 0.1), maximum tree depth (3, 4, 5), and number of estimators (200, 300, 400). Accuracy was used as the scoring metric for model selection. In addition, we restricted each tree to a random 80% of the training samples and 80% of the available features, a strategy to reduce overfitting by limiting reliance on any single subset of the data. The grid search identified the optimal configuration as a learning rate of 0.05, a maximum depth of 3, and 400 estimators. Our results are not sensitive to the choice of hyperparameters. Using an alternative prediction model, a fully connected neural network, achieves a very similar level of accuracy (see Web Appendix B).

Model Evaluation

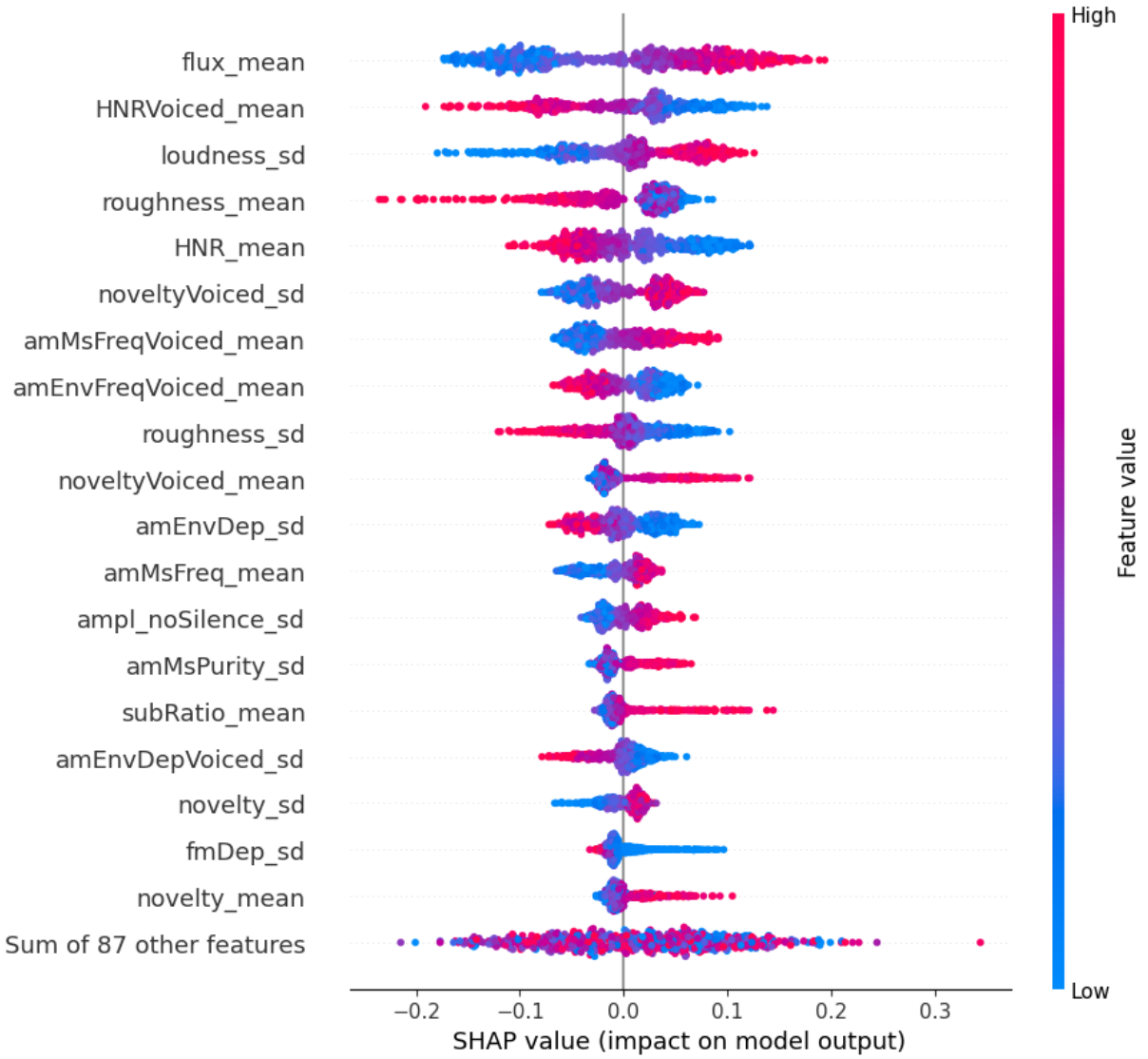
We start by evaluating the trained model using grouped five-fold cross-validation, which generated out-of-sample predictions while keeping all recordings from a given participant within the same fold. Model performance was assessed by comparing these predictions with the ground-truth labels. The classifier achieved an accuracy of 80.1% (95% CI: [.787, .813]), with an area under the curve (AUC) of .882. Sensitivity of 81.6% (95% CI: [.798., .833]) for correctly identifying strong preferences and specificity of 78.6% (95% CI: [.766, .804]) for correctly identifying weak preferences were similar.

To obtain an unbiased estimate of out-of-sample performance beyond cross-validation, we evaluated the model on a held-out test set that was collected independently using the same procedure as the training data and never seen during training (n = 623 observations). This evaluation is useful for assessing model generalization and making sure that the high accuracy seen before was not driven by model overfitting. Interestingly, the classifier achieved a slightly higher accuracy of 84.4% (95% CI: [.814, .871] and an AUC of .922. Sensitivity (86.5%, 95% CI: [.822, .899]) and a specificity of (82.4%, 95% CI: [.778, .862]) remained indistinguishable.

Together, these findings indicate that the model generalizes well to unseen data, providing further evidence of its robustness and limited susceptibility to overfitting.

When analyzing the SHAP values of the trained model, the results again highlight the role of spectral dynamism and voice quality in signaling preference strength (Figure 3). Spectral flux (i.e., the extent of rapid change in energy across frequencies) emerged as the single most influential predictor, reinforcing its central role as a marker of preference strength. Moreover, measures of voice quality, including the harmonic-to-noise ratio (HNR) (how "clean" or noise-free the voice sounds), also ranked among the strongest predictors, linking excessively pure voices to weaker preference strength. Lastly, loudness variability (i.e., moment-to-moment changes in amplitude) and roughness (i.e., subtle irregularities that add texture to the voice) also contributed meaningfully, underscoring that high preference strength are reflected in voices that are dynamic and textured rather than flat or uniform.

Figure 3: Summary of SHAP Values



Note: SHAP summary plot illustrating the impact of each feature on the model's predictions. Each dot represents a feature value for one observation. The position on the x-axis shows the SHAP value (effect on the model output), while color indicates the feature value (red = high, blue = low).

Discussion

While the machine learning model demonstrates that preference strength can be predicted from vocal paralinguistic cues with high accuracy, an important question remains: Do these predictions reflect patterns that humans can also perceive? To address this, we directly compared model performance with human judgment, asking whether listeners can detect preference strength from the same held-out test set used for model evaluation.

Study 2: Human Rater Validation

The main goal of Study 2 was to examine whether our machine learning model captures paralinguistic cues that humans naturally perceive or whether it detects speech artifacts beyond human sensitivity. We therefore recruited participants to judge consumers' preference strength from randomly sampled recordings of the hold-out test set used in the model evaluation stage to assess whether preference strength is indeed perceptible to human listeners.

Data Collection

We recruited 625 US-American participants via Prolific (M_{age} = 42.4, SD = 12.8; 51.8% female) whose first language is English to serve as human judges for preference strength detection.

After passing an audio test, participants listened to five audio recordings drawn from our test data (625 recordings from the held-out test data in the model evaluation), leading to a total of 3,125 ratings. For each recording, participants indicated whether they perceived the speaker's certainty in their decision as "strong" or "weak" and provided written explanations for their judgments. Lastly, participants reported their age and gender (see Web Appendix C for survey instrument).

Results and Discussion

As each participant rated a random sample of recordings from the full pool, the distribution of ratings across files was not uniform: although every file had an equal probability of being selected, some accumulated more ratings than others. To ensure reliable assessments and reduce dependence on individual rater idiosyncrasies, we restricted our analysis to the 409 voice recordings that received at least five independent ratings (total N = 2,442 ratings).

We then computed classification accuracy by comparing participants' ratings to ground-truth labels (i.e., the degree of preference strength speakers were instructed to convey). Human raters achieved an overall accuracy of 78.7% (95% CI: [.770, .803]) in detecting ground truth preference strength labels from vocal recordings. By contrast, our model reached an overall accuracy of 85.0% (95% CI: [.812, .884]), with a sensitivity of 87.7% (95% CI: [.825, .918]) for correctly identifying strong preferences and a specificity of 82.2% (95% CI: [.762, .873]) for correctly identifying weak preferences on the same test data.

As a robustness check, we repeated the analyses including all recordings regardless of rating count; results remained consistent. Specifically, human raters achieved an overall classification accuracy of 78.6% (95% CI: [.771, .800]) with a sensitivity of 80.9% (95% CI: [.789, .828]) and a specificity of 76.2% (95% CI: [.740, .783]).

To further deepen our understanding of what exactly our human raters picked up in the audio recordings, we analyzed their open-text responses across conditions using Hovy et al.'s (2021) Wordify text processing tool (see Web Appendix D for details). Raters' subjective descriptions aligned with the paralinguistic cues extracted by our model: terms such as "hesitant" or "uncertain" appeared more for weak preferences, whereas "confident," "decisive," or "quick" appeared more for strong preferences. This pattern is consistent with the model's feature-importance results—e.g., higher harmonic-to-noise ratio (clearer, more periodic voice) and greater energy/loudness are associated with stronger preferences.

The model's superior accuracy likely reflects its ability to capture fine-grained acoustic patterns that humans tend to summarize into higher-level semantic impressions (e.g., confident vs. uncertain, hesitant vs. quick/decisive) and may not consciously register. Thus, automated systems can replicate human intuitive judgments while also detecting subtle vocal cues of preference strength that operate below conscious awareness.

In sum, the performance of our machine learning model exceeds that of human judges, thereby validating our computational approach to inferring preference strength from vocal paralinguistic features. This alignment suggests that our trained XGBoost model effectively captures the complex vocal patterns that humans rely on when inferring preference strength from speech.

Model Applications

Study 3: Predicting Downstream Outcomes from Vocal Cues

To demonstrate the practical utility of our trained model, we applied it to a new dataset to obtain consumers' predicted preference strength using their voice data. Specifically, for each vocal expression, the model generated a probability that the speaker was conveying a strong preference. We then examined whether these probabilities predict participants' willingness to accept a price premium, thereby testing whether voice-based signals of preference—captured through our model—translate into behaviorally relevant outcomes. This application illustrates the broader value of our model for linking subtle paralinguistic cues to downstream consumer decisions.

Data Collection

We recruited 126 students for a lab study (M_{age} = 23.75, SD_{age} = 8.68; 71.6% female) participated in exchange for monetary compensation. The study employed a 12-replicate within-subjects design and consisted of two stages.

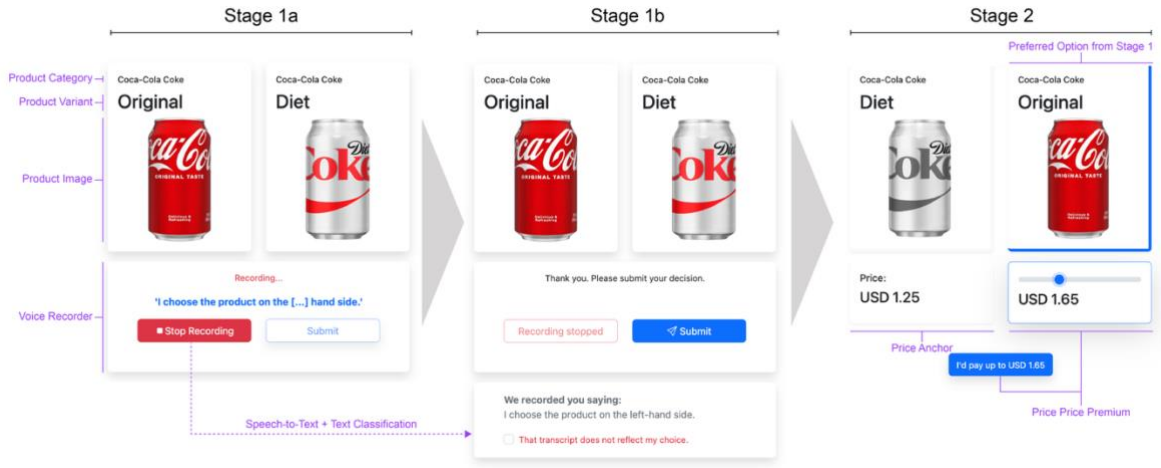
In stage 1a, participants completed twelve binary choice tasks between pairs of branded grocery products. Each trial displayed one product pair per page, and participants indicated their choice verbally. Responses were recorded through a microphone placed at their workstation, with

recording beginning automatically as soon as the product pair appeared. Each choice pair consisted of two variants from the same band (e.g., Ajax dish soap: lemon vs. orange; M&Ms: peanut vs. milk chocolate; Peet's Dark Roast Coffee: regular vs. decaf; see Web Appendix E for full list), thereby holding baseline prices constant. This ensured that decisions reflected preference strength rather than price sensitivity. Prices ranged from $1.25 to $9.95, covering a set of everyday grocery items that participants were highly likely to recognize, evaluate, and have established preferences for. The product pairs were determined based on a pretest (see below). The presentation order of product pairs and the left-right positioning of options within each pair were randomized for each participant. On every trial, participants vocalized their preferred product from the two alternatives. Importantly, to standardize semantic content, they were instructed to use the same sentence across all trials, only changing the respective side: *"I choose the product on the […] hand side."* In stage 1b, before submitting each choice, participants were shown a speech-to-text transcript of their preference and asked to confirm whether it accurately reflected what they had said; if not, they could repeat the previous step.

In stage 2, we measured price premium acceptance using an anchored price premium procedure consistent with prior research on relative willingness to pay (e.g., He, Anderson, and Rucker 2020; Steenkamp, van Heerde, and Geyskens 2010). Specifically, participants were reminded of their earlier choices from stage 1, along with the baseline prices of the unchosen alternatives. They then indicated the maximum price they were willing to pay for their chosen product by adjusting a slider ranging from 0% to 100% of the unchosen option's baseline price, in $0.05 increments (see Figure 4). This procedure anchors valuations to a reference product and thereby captures the relative price premium participants would accept for their preferred option. Lastly, participants reported their age and gender (see Web Appendix F for survey instrument).

We applied the same pre-processing procedures as in Study 1 but did not need to screen out any participant as all of them complied in the controlled lab environment.

Figure 4: Screenshot of the Custom Web Interface Built in oTree for Study 3



Pretest

We recruited 100 participants (M_{age} = 38.50, SD = 11.69; 47.0% female) from Prolific and asked them to indicate their preference strength across 18 pairs of product pairs by drawing a 101-point slider (-50 = Strong preference for left product, +50 = Strong preference for right product; similar to Nowlis and Simonson 1997). From this pretest, we selected 12 product pairs for the main study that showed the promise of producing relatively large variation in preference strength (see Web Appendices G and H).

Results and Discussion

To examine whether vocal expressions of preference strength predict participants' willingness to pay (WTP), we regressed the reported price premium (in percent of the unchosen option's baseline price) on the model-predicted probability of strong preference. In the simple OLS specification (Table 2, Model 1), higher predicted preference strength was associated with a significantly greater price premium ($\beta = 11.01$, SE = 1.87, $p < .001$). Since recent research has highlighted response duration—the interval between when the response begins (onset) and when

it ends (offset)—as a promising indicator of cognitive processing (Pfister, Neszmelyi, and Kunde 2023), we re-estimated the model with response duration as a covariate. The results remained robust and significant (β = 9.20, SE = 1.95, $p < .001$; Model 2).

Lastly, since we collected multiple price premium judgments for each participant, we further estimated a mixed-effects regression with random intercepts for participants. This is our preferred specification, as it emphasizes within-subject variation by examining how each participant's judgments deviate from their own baseline, while canceling out stable between-subject differences. The results were consistent with the OLS findings: model-predicted preference strength again significantly increased WTP (β = 11.25, SE = 2.05, $p < .001$; Model 3). Similarly, the results remained unchanged after controlling for response duration (β = 9.79, SE = 2.12, $p < .001$; Model 4). The variance component for participant intercepts was significant in both models, indicating systematic individual differences in baseline willingness to pay.

Table 2: Main Results Anchored Price Premium Acceptance

	Anchored Price Premium Acceptance in Percent											
	OLS						linear mixed-effects					
Model	(1)			(2)			(3)			(4)		
	b	*SE*	*p*	*b*	*SE*	*p*	*b*	*SE*	*p*	*b*	*SE*	*p*
Intercept	8.163	1.061	.000	12.610	1.751	.000	7.991	1.431	.000	11.597	1.992	.000
Model Prediction	11.009	1.870	.000	9.202	1.948	.000	11.251	2.046	.000	9.785	2.118	.000
Response Duration (seconds)				-1.070	0.336	.001				-0.866	0.334	.000
Participant Var							0.333	0.055	.000	0.329	0.055	.000
Observations	1,473			1,473			1,473			1,473		
R^2	0.023			0.030								
R^2 Adjusted	0.022			0.028								
Log Likelihood							−6,411.004			−6,407.659		

To illuminate vocal paralinguistic cues of preference strength that predict willingness to pay, we ran a LASSO regression (Tibshirani 1996) (Web Appendix I). Consistent with our preceding results, preference strength was systematically encoded in the spectral richness and amplitude patterns of a person's speech signal. Greater spectral entropy during speech (i.e., richer and more varied distribution of energy across frequencies) and higher harmonic richness (i.e., more organized and consistent frequency patterns in the voice) predicted higher preference strength, suggesting that more resonant and varied vocal qualities track with greater preference strength.

Together, these analyses demonstrate that paralinguistic cues captured by the voice-based model reliably predict participants' willingness to pay: vocal expressions of stronger preferences were associated with higher price premiums for the chosen product. Importantly, the model was originally trained in a different context (voice actors), yet it generalized effectively to predict willingness to pay in this new setting. This cross-context validity underscores the robustness of the model and highlights its potential for leveraging vocal paralinguistic signals to explain behaviorally relevant consumer outcomes.

Study 4: Predicting Preference Strength in Other Domains

The objectives of Study 4 were three-fold. First, we experimentally varied decisional conflict to induce differences in preference strength. Based on extant research (Festinger 1954; Tversky and Shafir 1992), when alternatives are close in utility, preferences should be weaker, whereas when one option clearly dominates, preferences should be stronger. This allowed us to validate that our model's predictions capture variation in preference strength driven by decisional conflict. Second, we aimed to extend our theorizing beyond product choices to an unrelated domain: social judgments. Research on impression formation suggests that social evaluations, such as judgments of attractiveness, also give rise to preferences that differ in relative utility and

strength (Fiske, Cuddy, and Glick 2007; Zajonc 1980). If our account is general, the same vocal paralinguistic cues that signal preference strength in consumer choices should emerge in this domain as well. Third, we sought to establish construct validity by examining whether model-inferred preference strength predicts participants' self-reported strength. Demonstrating convergence between the two provides theory-driven evidence that the vocal paralinguistic cues identified by our model capture a meaningful psychological construct.

Data Collection

We recruited 244 Prolific panelists to complete the study and applied the same exclusion criteria as in Study 1 which resulted in a final sample of 230 Prolific panelists (M_{age} = 39.54, SD = 13.08; 37.0% female). Our sample size was based on past research using within-subjects designs and a mouse-tracking task (Stillman and Ferguson 2020). The study employed a 2 (decisional conflict: low vs. high) × 2 (gender model: male vs. female) × 2 (replicates) within-subjects design.

Participants were asked to imagine that they were shortlisting models for an upcoming advertising campaign. Each participant indicated their choice verbally across eight binary choice trials, with a single pair of models presented per page. Four trials featured female pairs (two low-conflict, two high-conflict) and four featured male pairs (two low-conflict, two high-conflict). All pairs were selected based on a pretest (see below) and the order of trials and left-right placement of models was randomized for each participant.

To standardize verbal content, participants were instructed to use the same sentence across all trials, only changing the respective letter: *"I choose Model X [Y]."* Two practice rounds familiarized participants with the interface before the main task. To accurately capture response times, participants were informed that the recording would start automatically for each choice trial and that they had to stop the recording once they had expressed their choice.

After each choice, participants reported their preference strength on an unnumbered 101-point slider scale (i.e., -50 = Strong preference for Model X, +50 = strong preference for Model Y). Lastly, participants indicated their age and gender (see Web Appendix J for survey instrument).

Pretest

We conducted two separate pretests to identify low- and high-conflict stimuli for the main study. In one pretest, we recruited 131 participants (M_{age} = 24.28, SD = 2.76; 40.5% female) to rate nine pairs of female models who appeared similarly attractive (matched pairs, intended to create high decisional conflict) and nine pairs of female models who differed clearly in attractiveness (mismatched pairs, intended to induce low decisional conflict; see Web Appendices K and L). Similarly, we recruited 100 participants (M_{age} = 39.72, SD = 14.24; 43.0% female) to rate six pairs of male models who appeared similarly attractive and five pairs of male models who differed clearly in attractiveness (see Web Appendices M and N). This procedure followed previous research that manipulates decisional conflict by varying the magnitude of attractiveness between options (Dhar 1997; Miller 1944; Ülkümen and Malkoc 2015). From this pretest, we selected, for each gender, the two pairs that induced the lowest decisional conflict (mismatched in attractiveness) and the two pairs that induced the highest decisional conflict (matched in attractiveness).

Results and Discussion

Consistent with prior research on decisional conflict (e.g., Dhar 1997; Ülkümen and Malkoc 2015), a repeated-measures ANOVA with decisional conflict (low vs. high) as the independent variable and self-reported preference strength as the dependent variable suggests that participants reported weaker preferences in high- compared to low-conflict trials (M_{low} = 45.04,

SD = 8.87 vs. M_{high} = 19.86, SD = 16.00; F(1, 229) = 838.53, $p < .001$). This confirms that our manipulation successfully varied decisional conflict as intended.

Next, we examined whether the model-predicted probability of strong preference predicts participants' self-reported preference strength. An OLS regression revealed a significant positive effect of the model prediction (β = 6.14, SE = 1.51, $p < .001$; Table 3, Model 1), indicating that higher predicted probabilities corresponded to stronger reported preferences. The results remained robust and significant after controlling for response duration (β = 5.18, SE = 1.54, $p < .001$; Model 2).

To account for the repeated measurements within participants, we further estimated a mixed-effects regression with random intercepts for participants. The results were consistent with the OLS findings: model-predicted preference probability significantly predicted self-reported preference strength (β = 7.97, SE = 1.68, $p < .001$; Model 3). The results remained robust and significant after controlling for response duration (β = 6.96, SE = 1.72, $p < .001$; Model 4). The variance component for participant intercepts was significant in both mixed-effects models, indicating systematic individual differences in baseline self-reported preference strength.

Table 3: Main Results Self-Reported Preference Strength

Model	Self-Reported Preference Strength (50-point)											
	OLS						linear mixed-effects					
	(1)			(2)			(3)			(4)		
	b	*SE*	*p*	*b*	*SE*	*p*	*b*	*SE*	*p*	*b*	*SE*	*p*
Intercept	30.602	0.618	.000	33.102	1.021	.000	30.052	0.719	.000	34.462	1.189	.000
Model Prediction	6.142	1.511	.000	5.179	1.540	.001	7.971	1.682	.000	6.962	1.722	.000
Response Duration (seconds)				-0.769	0.250	.002				-1.429	0.302	.000
Participant Var							0.074	0.020	.000	0.110	0.026	.000
Observations	1,840			1,840			1,840			1,840		

R^2	0.009	0.014		
R^2 Adjusted	0.008	0.013		
Log Likelihood			−7,913.588	−7,901.845

Note: * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$

As in the preceding study, a LASSO regression further underscored the importance of variability- and quality-related vocal paralinguistic features (see Web Appendix O). Greater overall variability in amplitude (i.e., moment-to-moment fluctuations in loudness) and spectral flux (i.e., the degree of rapid change in spectral energy from one moment to the next) were positively associated with preference strength, reinforcing that dynamic and energetic voices encode stronger preferences.

Together, these results provide additional validation for our theorized mechanism: by experimentally manipulating decisional conflict—an antecedent of preference strength—we showed that the model's predictions reliably tracked differences in self-reported preference strength. This strengthens the claim that the vocal paralinguistic features captured by our model are not merely correlates of preference strength but valid indicators of its underlying mechanism. Furthermore, although this study asked participants to choose between advertising models, a very different task from the willingness to pay paradigm used in Study 3, the converging results underscore the model's broad applicability across distinct decision domains.

General Discussion

Marketing research has long relied on measures of preference strength, such as people's written self-reports, response times, or mouse-tracking. Some of these established measures come with important shortcomings in their application to marketplace settings. In line with recent calls for new data sources and modality-specific frameworks (e.g., Melzner and Raghubir 2023), VIPS

departs from previous work by leveraging vocal paralanguage to assess preference strength. Across four studies, we develop a controlled, experimentally labeled dataset, train machine learning models on vocal paralinguistic features (Study 1), and validate the approach with human raters who achieve comparable accuracy (Study 2). Finally, we demonstrate that voice-derived measures of preference strength extend to behaviorally relevant consumer outcomes, including willingness to pay (Study 3) and self-reported preference strength (Study 4), underscoring the potential of vocal paralanguage as a scalable and unobtrusive tool for researchers and firms.

Theoretical Implications

This research makes four key contributions to the marketing and decision-making literature. First, we extend the literature on vocal paralanguage by shifting the focus from its influence on receivers to its diagnostic value reflecting a critical decision state of a speaker (their momentary preference strength). Prior studies have demonstrated that vocal cues influence persuasion (Xu et al. 2025), product perceptions (Lowe and Haws 2017), or funding decisions (Wang et al. 2021), yet little is known about whether the subtleties of *how* something is said reflect the speaker's own underlying state. By leveraging a machine learning model trained on vocal paralinguistic cues, this research positions vocal paralanguage as a real-time, largely involuntary signal of consumers' cognitive and affective processes during a decision.

Second, we contribute to the literature on preference strength measurement by introducing vocal paralanguage as a novel and naturally occurring indicator. Traditional approaches—such as self-reports, text analysis, response times, or mouse-tracking—all come with important shortcomings. For example, response time captures nondecision time and/or is difficult to interpret due to its nonlinear relationship with preference strength (Evans, Dillon, and Rand 2015; Luce 1986). More recent methods like mouse-tracking capture decisional conflict dynamically but are impractical in touch- and voice-driven market contexts. Our findings identify

that variability-related features (e.g., spectral and amplitude variability) and voice-quality markers (e.g., harmonic-to-noise ratios) covary systematically with preference strength. This extends the measurement literature by offering a continuous, unobtrusive, and ecologically valid indicator of consumers' confidence in their choices.

Third, we add to the literature on decisional conflict and its behavioral consequences by demonstrating that vocal paralinguistic cues not only mirror internal decisional conflict but also predict behaviorally relevant outcomes. Specifically, we show that voice-derived preference strength forecasts willingness to pay, linking real-time decision dynamics to economically meaningful behavior. In doing so, our research enriches theories of decision-making by illuminating how the moment-to-moment resolution of decisional conflict is embedded in the human voice and ultimately translates into actual market actions.

Practical Implications

Our research offers actionable guidance for managers on how to leverage vocal paralinguistic cues as a window into consumers' underlying preference strength as they make a choice. Given the rapid rise of voice-first commerce and the ubiquity of "vocalized interactions"—whether through call centers, smart speakers, mobile voice assistants, or drive-thru ordering systems—firms are increasingly in a position to analyze not only *what* consumers say, but also *how they sound* when say it. Our findings suggest that vocal paralanguage provides a scalable and unobtrusive indicator of preference strength at the moment of choice. This has several implications: firms can improve selection by identifying when consumers are uncertain and may require assistance; they can tailor promotions or adjust pricing strategies based on the likelihood of accepting a price premium; and they can personalize customer experiences by adapting recommendations or support levels to the confidence with which consumers express preferences. Importantly, because vocal paralanguage is expressed naturally, without the need for

additional questions or interruptions, these insights can be gathered seamlessly, reducing friction and enhancing the customer journey. More broadly, our approach opens opportunities for firms to integrate real-time, voice-based analytics into customer relationship management systems, enabling smarter, context-sensitive decisions in environments where timing and personalization are critical.

Policy Implications and Ethical Considerations

The ability to infer preference strength from vocal paralinguistic features raises important policy considerations, particularly because voice interfaces already routinely record and analyze what consumers say. Our approach extends this analysis to *how* they are said, revealing psychological states without consumers' explicit awareness. This distinction matters for informed consent: while consumers may understand their words are transcribed, they are unlikely to realize their vocal tone and hesitations can reveal information about their underlying preference strength. Such capabilities could create risks of exploiting consumer vulnerability, as firms might detect moments of maximum uncertainty to deploy persuasive interventions, or identify markers of strong preference to implement dynamic pricing strategies that extract higher prices.

From a data governance perspective, vocal paralinguistic analysis transforms routine voice interactions into rich psychological profiles, raising questions about data minimization, purpose limitation, and retention periods. Current regulatory frameworks like GDPR and CCPA may apply to voice data as personal, and potentially as biometric information, yet they do not specifically address inferring subtle psychological states from vocal paralinguistic features. We recommend that firms adopting these technologies implement transparent disclosure practices, obtain explicit consent for vocal paralinguistic analysis beyond transcription, enforce strict use limitations that prohibit exploitative practices, and establish internal review processes to ensure that applications serve consumer welfare rather than undermining it. As voice interfaces become

increasingly embedded in commerce, proactive self-regulation by industry may be preferable to reactive regulatory intervention.

Technical Contributions

Beyond managerial applications, our contribution also extends to practice by offering a unique, experimentally labeled dataset of vocalized preferences—a resource that directly answers the call for new methods and benchmarks in marketing research. Importantly, our custom web interface also provides a practical contribution in its own right, and offers researchers a scalable and adaptable platform for collecting high-quality, voice-based behavioral data in controlled experimental settings. This dataset and infrastructure provide a benchmark for developing and testing new methods, supporting replication, and fostering innovation at the intersection of marketing, decision-making, and computational linguistics. More broadly, our approach opens opportunities for firms to integrate real-time, voice-based analytics into customer relationship management systems, enabling smarter, context-sensitive decisions in environments where timing and personalization are critical.

In addition, we provide practical resources to support future research and applications. We developed an open-source, web-based Shiny application (available at https://vipstest.shinyapps.io/shiny-app/) that allows researchers and practitioners to upload their speech data and generate predictions of preference strength. This tool makes it easy to explore our approach without requiring extensive technical infrastructure. Together, these resources lower the barriers to uilizing speech data in marketing.

Limitations & Future Research

Like all research, ours also comes with limitations. Our training and validation data were recorded on people's own devices in everyday settings rather than in a tightly controlled lab. That choice can lower accuracy compared with lab-quality recordings (Busquet, Efthymiou, and

Hildebrand 2024). However, we made this decision on purpose so the data would look and sound like the real situations where this tool would be used. Also, our system runs in near-time. Given how lightweight our technical setup is—and how quickly speech and NLP tools are improving—true real-time use seems within reach, but it will take additional engineering and testing.

Future research may also further validate our findings in field settings and across diverse channels (e.g., call centers, smart speakers, drive-thru contexts), devices, cultures, languages, and accents to ensure robustness and generalizability. Building multilingual models will help here, because people express themselves differently across languages. We also want to tie voice cues to real behavior by using tasks with actual money at stake, so we can examine whether the same vocal paralinguistic patterns that signal preference strength also predict real purchasing decisions. Another goal is to track people over time to see if these voice markers are stable for the same person or if they change with context, mood, or practice. Because real environments are noisy, we need to stress-test the system with background sounds—despite modern noise filtering—to make sure performance stays strong. Finally, combining voice with other simple signals (e.g., body language, or basic information from wearables) could make predictions more accurate and easier to explain.

References

Anderson, Eric T. and Duncan I. Simester (2014), "Reviews Without a Purchase: Low Ratings, Loyal Customers, and Deception," *Journal of Marketing Research*, 51 (3), 249–69.

Bettman, James R. and Mita Sujan (1987), "Effects of framing on evaluation of comparable and noncomparable alternatives by expert and novice consumers," *Journal of Consumer Research*, 14 (2), 141–54.

Busquet, Francesc, Fotis Efthymiou, and Christian Hildebrand (2024), "Voice analytics in the wild: Validity and predictive accuracy of common audio-recording devices," *Behavior Research Methods*, 56 (3), 2114–34.

Busquet, Francesc and Christian Hildebrand (2023), voiceR: Voice Analytics for Social Scientists. https://CRAN.R-project.org/package=voiceR.

Chen, Daniel L., Martin Schonger, and Chris Wickens (2016), "oTree—an Open-Source Platform for Laboratory, Online, and Field Experiments," *Journal of Behavioral and Experimental Finance*, 9, 88–97.

Chernev, Alexander (2001), "The impact of common features on consumer preferences: A case of confirmatory reasoning," *Journal of Consumer Research*, 27 (4), 475–88.

Cittadini, Riccardo, Chiara Tamantini, Francesco Scotto di Luzio, Cristiano Lauretti, Luca Zollo, and Francesco Cordella (2023), "Affective state estimation based on Russell's model and physiological measurements," *Scientific Reports*, 13 (1), 9786.

Cummins, Nicholas, Alice Baird, and Björn W. Schuller (2018), "Speech Analysis for Health: Current State-of-the-Art and the Increasing Impact of Deep Learning," *Methods*, 151, 41–54.

Darley, James (2025), "How McDonald's is Integrating AI Across Restaurants Globally," Technology Magazine, March. https://technologymagazine.com/articles/how-mcdonalds-is-using-ai-to-boost-efficiency-and-workflows. Accessed: September 15, 2025.

Dhar, Ravi (1997), "Consumer preference for a no-choice option," *Journal of Consumer Research*, 24 (2), 215–31.

Evans, Anthony M., Kyle D. Dillon, and David G. Rand (2015), "Fast but not intuitive, slow but not reflective: Decision conflict drives reaction times in social dilemmas," *Journal of Experimental Psychology: General*, 144 (5), 951–66.

Festinger, Leon (1954), "The Social Comparison Theory," *Human Relations*, 7, 117–40.

Fisher, Geoffrey and Kaitlin Woolley (2024), "How consumers resolve conflict over branded products: Evidence from mouse cursor trajectories," *Journal of Marketing Research*, 61 (1), 165–84.

Fiske, Susan, Amy J. Cuddy, and Peter Glick (2007), "Universal dimensions of social cognition: Warmth and competence," *Trends in Cognitive Sciences*, 11 (2), 77–83.

Freeman, Jonathan B. (2018), "Doing psychological science by hand," Current Directions in *Psychological Science*, 27 (5), 315–23.

Gandomi, Amir and Murtaza Haider (2015), "Beyond the hype: Big data concepts, methods, and analytics," *International Journal of Information Management*, 35 (2), 137–44.

Grand View Research (2024), "Voice Commerce Market Size, Share & Trends Analysis Report By Device Type (Smart Speakers, Smartphones), By Industry Vertical (Consumer Goods & Retail, Healthcare, Automotive), By Region, And Segment Forecasts, 2024 - 2030," Market Research Report, Grand View Research. https://www.grandviewresearch.com/industry-analysis/voice-commerce-market-report. Accessed: September 15, 2025.

Hancock, Jeffrey T., Lauren E. Curry, Saurabh Goorha, and Michael Woodworth (2007), "On Lying and Being Lied To: A Linguistic Analysis of Deception in Computer-Mediated Communication," *Discourse Processes*, 45 (1), 1–23.

He, Sharlene, Eric T. Anderson, and Derek D. Rucker (2024), "Measuring willingness to pay: A comparative method of valuation," *Journal of Marketing*, 88 (3), 50–68.

Hovy, Dirk, Shiri Melumad, and Jeff Inman (2021), "Wordify: A tool for discovering and differentiating consumer vocabularies," *Journal of Consumer Research*, 48 (3), 394–414.

Jurafsky, Dan and James H. Martin (2021). *Speech and language processing* (Vol. 3). Prentice-Hall.

Khan, Uzma and Ravi Dhar (2006), "Licensing effect in consumer choice," *Journal of Marketing Research*, 43 (2), 259–66.

Kleiman, Tali and Ran R. Hassin (2011), "Non-conscious goal conflicts*," Journal of Experimental Social Psychology*, 47 (3), 521–32.

Konovalov, Arkady and Ian Krajbich (2019), "Revealed strength of preference: Inference from response times," *Judgment and Decision Making*, 14 (4), 381–94.

Ligges, Uwe, Sebastian Krey, Olaf Mersmann, and Sarah Schnackenberg (2013), tuneR: Analysis of music. *URL: http://r-forge. r-project. org/projects/tuner*.

Loomes, Graham (2005), "Modelling the stochastic component of behaviour in experiments: Some issues for the interpretation of data," *Experimental Economics*, 8 (4), 301–23.

Lowe, Michael L. and Kaitlin L. Haws (2017), "Sounds big: The effects of acoustic pitch on product perceptions," *Journal of Marketing Research*, 54 (2), 331–46.

Luce, R. Duncan (1986), *Response Times: Their Role in Inferring Elementary Mental Organization*. Oxford, UK: Oxford University Press.

Ludwig, Stephan, Tom Van Laer, Ko De Ruyter, and Mike Friedman (2016), "Untangling a Web of Lies: Exploring Automated Detection of Deception in Computer-Mediated Communication," *Journal of Management Information Systems*, 33 (2), 511–41.

Malkoc, Selin A., Gal Zauberman, and Cenk Ulu (2005), "Consuming now or later? The interactive effect of timing and attribute alignability," *Psychological Science*, 16 (5), 411–7.

Market.us (2025), "Global Voice Commerce Market Size, Share, Statistics Analysis Report By Device Type (Smart Speakers, Smartphones), By Industry Vertical (Consumer Goods & Retail, Healthcare, Automotive, Travel & Tourism, Others), Region and Companies: Industry Segment Outlook, Market Assessment, Competition Scenario, Trends and Forecast 2025-2034," Market Research Report, Market.us. https://market.us/report/voice-commerce-market/. Accessed: September 15, 2025.

Melzner, Johann, Andrea Bonezzi, and Tom Meyvis (2023), "Information disclosure in the era of voice technology," *Journal of Marketing*, 87 (4), 491–509.

Melzner, Johann and Priya Raghubir (2023), "The sound of music: The effect of timbral sound quality in audio logos on brand personality perception," *Journal of Marketing Research*, 60 (5), 932–49.

Mira, Lea (2024), "Taco Bell Embraces AI Drive-Thrus, Aiming for Increased Restaurant Efficiency and Customer Satisfaction," Restaurant Technology News, August. https://restauranttechnologynews.com/2024/08/taco-bell-embraces-ai-drive-thrus-aiming-for-increased-restaurant-efficiency-and-customer-satisfaction/. Accessed: September 15, 2025.

Mogilner, Cassie, Jennifer Aaker, and Sepandar D. Kamvar (2012), "How happiness affects choice," *Journal of Consumer Research*, 39 (2), 429–43.

Moon, Sangkil and Wagner A. Kamakura (2017), "A picture is worth a thousand words: Translating product reviews into a product positioning map," *International Journal of Research in Marketing*, 34 (1), 265–285.

Netzer, Oded, Alain Lemaire, and Michal Herzenstein (2019), "When Words Sweat: Identifying Signals for Loan Default in the Text of Loan Applications," *Journal of Marketing Research*, 56 (6), 960–80.

Nisbett, Richard E. and Timothy D. Wilson (1977), "Telling more than we can know: Verbal reports on mental processes," *Psychological Review*, 84 (3), 231–259.

Nowlis, Stephen M. and Itamar Simonson (1997), "Attribute–task compatibility as a determinant of consumer preference reversals," *Journal of Marketing Research*, 34 (2), 205–18.

Ott, Myle, Claire Cardie, and Jeffrey Hancock (2012), "Estimating the Prevalence of Deception in Online Review Communities," in *Proceedings of the 21st International Conference on World Wide Web*. New York: Association for Computing Machinery, 201–10.

Pfister, Roland, Bálint Neszmélyi, and Wilfried Kunde (2023), "Response durations: A flexible, no-cost tool for psychological science," *Current Directions in Psychological Science*, 32 (2), 160–6.

Pham, Michel Tuan (2013), "The seven sins of consumer psychology," *Journal of Consumer Psychology*, 23(4), 411–23.

Pham, Michel Tuan, Aradhna Faraji-Rad, Olivier Toubia, and Leonard Lee (2015), "Affect as an ordinal system of utility assessment," *Organizational Behavior and Human Decision Processes*, 131, 81–94.

Podsakoff, Philip M., Scott B. MacKenzie, Jeong-Yeon Lee, and Nathan P. Podsakoff (2003), "Common method biases in behavioral research: a critical review of the literature and recommended remedies," *Journal of Applied Psychology*, 88 (5), 879–903.

Ratcliff, Roger, Philip L. Smith, Scott D. Brown, and Gail McKoon (2016), "Diffusion Decision Model: Current Issues and History," *Trends in Cognitive Sciences*, 20 (4), 260–81.

Rocklage, Matthew D., Sharlene He, Derek D. Rucker, and Loran F. Nordgren (2023), "Beyond sentiment: The value and measurement of consumer certainty in language," *Journal of Marketing Research*, 60 (5), 870–88.

Rude, Stephanie, Eva-Maria Gortner, and James Pennebaker (2004), "Language Use of Depressed and Depression-Vulnerable College Students," *Cognition & Emotion*, 18 (8), 1121–33.

Scherer, Klaus R. (1986), "Vocal affect expression: a review and a model for future research," *Psychological Bulletin*, 99 (2), 143–65.

Scherer, Klaus R., Tom Johnstone, and Gudrun Klasmeyer (2003), "Vocal Expression of Emotion," in *Handbook of Affective Sciences*, Richard J. Davidson, Klaus R. Scherer, and H. Hill Goldsmith, eds. Oxford University Press, 433–56.

Schuller, Björn and Anton Batliner (2013), *Computational Paralinguistics: Emotion, Affect and Personality in Speech and Language Processing*. John Wiley & Sons.

Song, Xiaobing, Jihye Jung, and Yinlong Zhang (2021), "Consumers' preference for user-designed versus designer-designed products: The moderating role of power distance belief," *Journal of Marketing Research*, 58 (1), 163–81.

Song, Joo-Hyun and Ken Nakayama (2009), "Hidden Cognitive States Revealed in Choice Reaching Tasks," *Trends in Cognitive Sciences*, 13 (8), 360–66.

Spiliopoulos, Leonidas and Andreas Ortmann (2018), "The BCD of response time analysis in experimental economics," *Experimental Economics*, 21 (2), 383–433.

Spiller, Stephen A. and Lyudmila Belogolova (2017), "On consumer beliefs about quality and taste," *Journal of Consumer Research*, 43 (6), 970–91.

Spivey, Michael (2008), *The continuity of mind*. Oxford University Press.

Steenkamp, Jan-Benedict E. M., Harald J. Van Heerde, and Inge Geyskens (2010), "What makes consumers willing to pay a price premium for national brands over private labels?," *Journal of Marketing Research*, 47 (6), 1011–24.

Stillman, Paul E. and Melissa J. Ferguson (2019), "Decisional Conflict Predicts Impatience," *Journal of the Association for Consumer Research*, 4 (1), 47–56.

Stillman, Paul E., Xianyun Shen, and Melissa J. Ferguson (2018), "How mouse-tracking can advance social cognitive theory," *Trends in Cognitive Sciences*, 22 (6), 531–543.

Tibshirani, Robert (1996), "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 58 (1), 267–88.

Tversky, Amos and Eldar Shafir (1992), "Choice under conflict: The dynamics of deferred decision," *Psychological Science*, 3 (6), 358–61.

Ülkümen, Gülden and Selin A. Malkoc. *Category Width Influences Sensitivity to Comparison Frames*. [unpublished manuscript]

Van Zant, Alex B. and Jonah Berger (2020), "How the voice persuades," *Journal of Personality and Social Psychology*, 118 (4), 661–82.

Wang, Xin, Shijie Lu, Xiaolin Li, Mansur Khamitov, and Neil Bendle (2021), "Audio Mining: The Role of Vocal Tone in Persuasion," *Journal of Consumer Research*, 48 (2), 189–211.

Xu, Haiyan, Yiming Ding, Yujia Ding, Qian Zhang, and Chenting Zhang (in press), "Whispers in your mind: the role of voice features in customer acquisition and retention," *Journal of Marketing*.

Yoon, So-Young O. and Itamar Simonson (2008), "Choice set configuration as a determinant of preference attribution and strength," *Journal of Consumer Research*, 35 (2), 324–36.

Zajonc, Robert B. (1980), "Feeling and thinking: Preferences need no inferences," *American Psychologist*, 35 (2), 151–75.

Zierau, Nils, Christian Hildebrand, Andreas Bergner, Francesc Busquet, Alexander Schmitt, and Jan Marco Leimeister (2023), "Voice bots on the frontline: Voice-based interfaces enhance flow-like consumer experiences & boost service outcomes," *Journal of the Academy of Marketing Science*, 51 (4), 823–42.

Web Appendix

Web Appendix A. Survey instrument (Study 1)

Consent

Welcome!

Thank you for your interest in this study.

Institution: XXX

Technical Requirements: We will ask you to provide a voice input. For that reason, we require you to use a device with a microphone in a quiet and private place. For the smoothest possible process, we advise using a modern browser other than Safari with a stable Internet connection.

☐ I confirm that I am using a sufficiently good microphone (e.g., an USB microphone) to complete the study.

☐ I confirm that I complete the study in a silent environment without any background noises (e.g., traffic or music).

Duration: Approximately 4 minutes.

Privacy:

☐ I agree to the anonymized analysis of my answers for the purpose of scientific evaluation according to the privacy policy.

[PAGE BREAK]

Instructions I

On the next pages, you will be asked to imagine choosing between two products or services and saying your decision out loud.

In particular, your task is to read two given decisions out loud, each in three different styles. Importantly, you cannot change the words you say, but you should choose to say them in the following three styles:

NEUTRAL

Here, your goal is to say the given decision aloud as you normally would.

STRONG PREFERENCE / WITH CERTAINTY

Here, your goal is to read the given decision out loud in a way that conveys a strong preference and certainty when choosing. Express that you are absolutely certain about your choice because it is easy for you to tell which option stands out as better.

WEAK PREFERENCE / WITH UNCERTAINTY

Here, your goal is to read the given decision out loud in a way that conveys a weak preference and uncertainty when choosing. Express that you are uncertain about your choice because it is hard for you to tell which option stands out as better.

You will need to allow your browser to access your microphone to proceed with this study. Please click the *Allow microphone recordings* button below when you're ready to continue.

[PAGE BREAK]

Instructions II

Before you begin

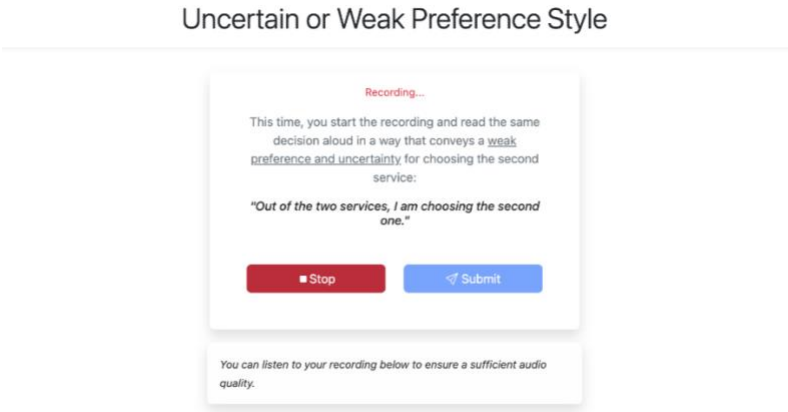
Please ensure you are in a quiet environment with no background noise (e.g., open windows with traffic noise, music playing, or conversations of others).

Please read all instructions carefully before recording.

[PAGE BREAK]

Task (2 sentences, 3 preference styles (neutral, weak, strong)

Uncertain or Weak Preference Style (exemplary))



[PAGE BREAK]

Demographics

- Please enter your age:
- Please select your gender. (1 = Female, 2 = Male, 3 = Other, 4 = Prefer not to say)

[PAGE BREAK]

Please use the box below to share your thoughts about the study (e.g., any aspects you particularly liked or disliked) to help us improve it.

Web Appendix B. Neural Network Results (Study 1)

The main analysis restricted the dataset to voice recordings that received at least 5 independent ratings to ensure reliable human judgments and minimize individual rater idiosyncrasies. While this approach attempts to improve measurement quality, it reduces the available sample size from 620 to 409 recordings. To demonstrate that our conclusions about human-model performance comparisons remain robust across different analytical choices, we present a parallel analysis that includes all voice recordings.

Using the same methodology described in the main text, we analyzed the complete set of 620 voice recordings from the test data that received human ratings. Each recording was classified as expressing either “strong” or “weak” preference strength based on human judgments, and these classifications were compared against ground truth labels and machine learning model predictions.

Human raters achieved an overall classification accuracy of 78.6%% (95% CI: [.771, .800]) on the full dataset, with a sensitivity of .809 for correctly identifying strong preferences and a specificity of .762 for correctly identifying weak preferences. These values closely mirror the main analysis results of 78.7%% overall accuracy, .807% sensitivity, and .766% specificity.

Web Appendix C. Survey instrument (Study 2)

Consent

Welcome!

Thank you for your interest in this study. It is anonymous and there are no right or wrong answers.

Study: Classification Task

Consent:

☐ I agree to the anonymized analysis of my answers for the purpose of scientific evaluation according to the consent form.

[PAGE BREAK]

Instructions

You will listen to short audio recordings of people speaking and rate how certain each speaker sounds about their decision.

Task overview:

First, you'll complete a brief control question with an audio sample

Then you'll rate a few audio recordings on the speaker's certainty

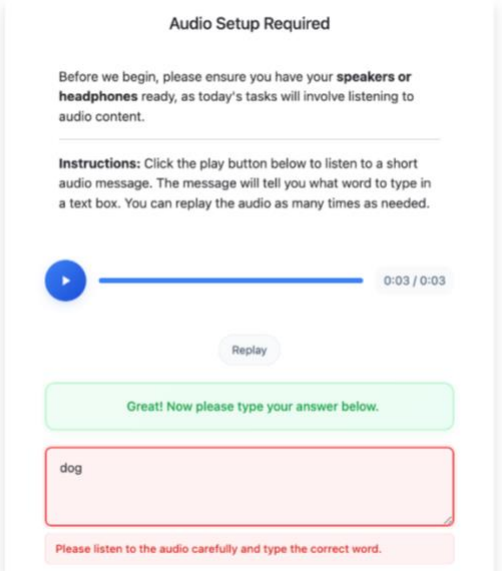
For each recording, you can replay the audio as many times as needed

Finally, a few brief follow-up questions

Click "Continue" when you're ready to begin with the control question.

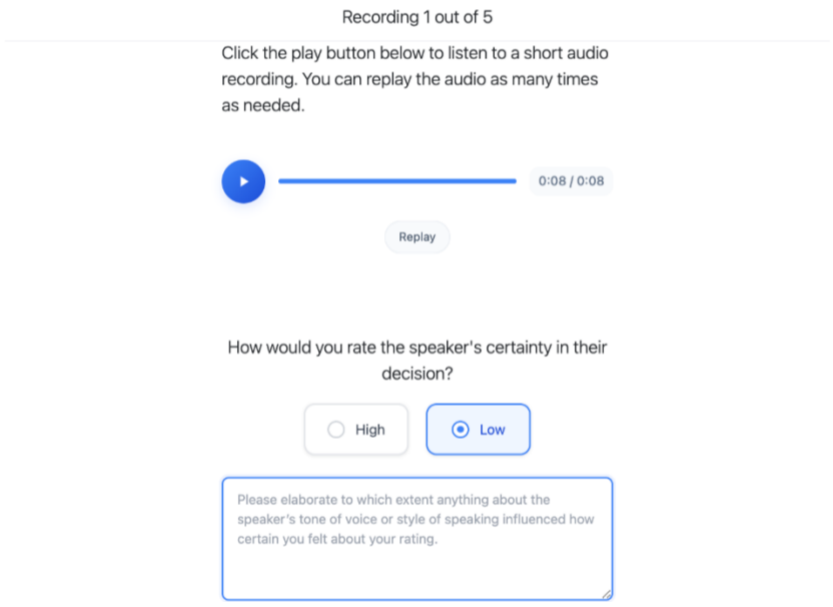
[PAGE BREAK]

Audio Setup



[PAGE BREAK]

Task (5 recordings, order randomized)



[PAGE BREAK]

Demographics

- Please enter your age:
- Please select your gender. (1 = Female, 2 = Male, 3 = Other, 4 = Prefer not to say)

Web Appendix D. Wordify Results (Study 2)

The Wordify analysis revealed that raters' subjective descriptions aligned with the paralinguistic cues extracted by our model. Words such as hesitant and unsure were more frequently associated with weak preferences, whereas words such as confident, decisive, and quick were more often used to describe strong preferences. Importantly, each of these lexical items also showed the opposite association when applied to the other condition, suggesting that they reliably discriminate between weak and strong preference expressions. The strongest lexical marker was confident, which exhibited the highest positive association with strong preferences and the strongest negative association with weak preferences. This lexical pattern is consistent with the model's feature-importance results—e.g., higher harmonic-to-noise ratio (clearer, more periodic voice) and greater energy/loudness are associated with stronger preferences.

Table: Wordify Results

	Word	**Score**	**Label**	**Correlation**
0	Hesitant	0.316	Low	Positive
1	Unsure	0.29	Low	positive
2	Like	0.282	Low	Positive
3	Confident	0.386	High	positive
4	Decisive	0.31	High	Positive
5	Quick	0.294	High	Positive
6	Decision	0.252	High	Positive
7	Confident	0.386	Low	Negative
8	Decisive	0.31	Low	Negative

9	Quick	0.294	Low	Negative
10	Decision	0.252	Low	Negative
11	Hesitant	0.316	High	Negative
12	Unsure	0.29	High	Negative
13	like	0.282	high	Negative

Web Appendix E. Final Stimuli (Study 3)

$3.35

$1.25

$6.95

$5.55

$1.35

$3.95

$1.75

$1.25

$9.95

$3.45

$4.45

$4.95

Web Appendix F. Survey Instrument (Study 3)

Consent

Welcome!

Thank you for your interest in this study.

Study: Decision articulation study

Institution: XXXX

Duration: Approximately 10 minutes.

Technical Requirements: We will ask you to provide a voice input. For that reason, we require you to use a device with a microphone in a quiet and private place. For the smoothest possible process, we advise using a modern browser other than Safari with a stable Internet connection.

Privacy:
☐ I agree to the anonymized analysis of my answers for the purpose of scientific evaluation according to the privacy policy.

[PAGE BREAK]

Instructions

This study consists of four parts: one set of trial rounds, two stages in which you make choices, and one short questionnaire.

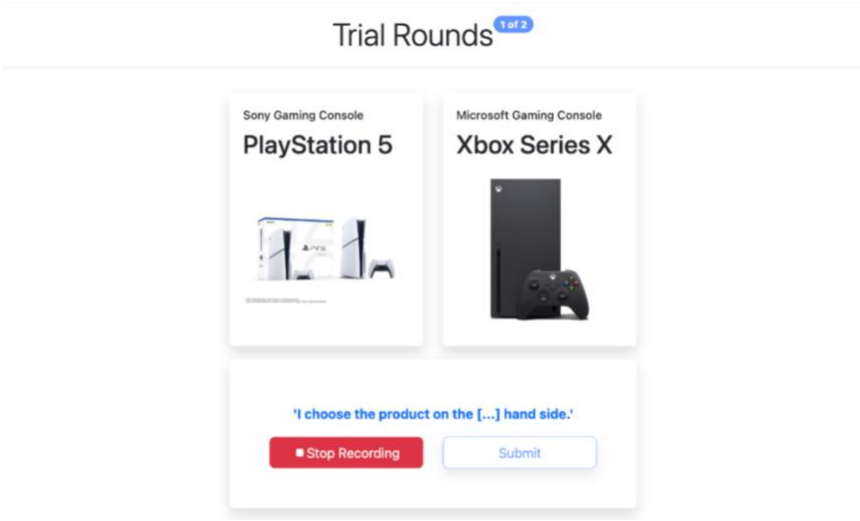
You will start with 2 trial rounds to get familiar with the interface and audio recording.
In each trial round, you'll see a set of two products. Choose your preferred product by saying, "I choose the product on the [left/right] hand side" aloud.

The interface will automatically start recording your response, and you can listen to it afterward. Adjust settings and re-record as needed to ensure clear audio without background noise.

Click the button below to allow microphone access. Once granted, you can proceed with the two trial rounds.

[PAGE BREAK]

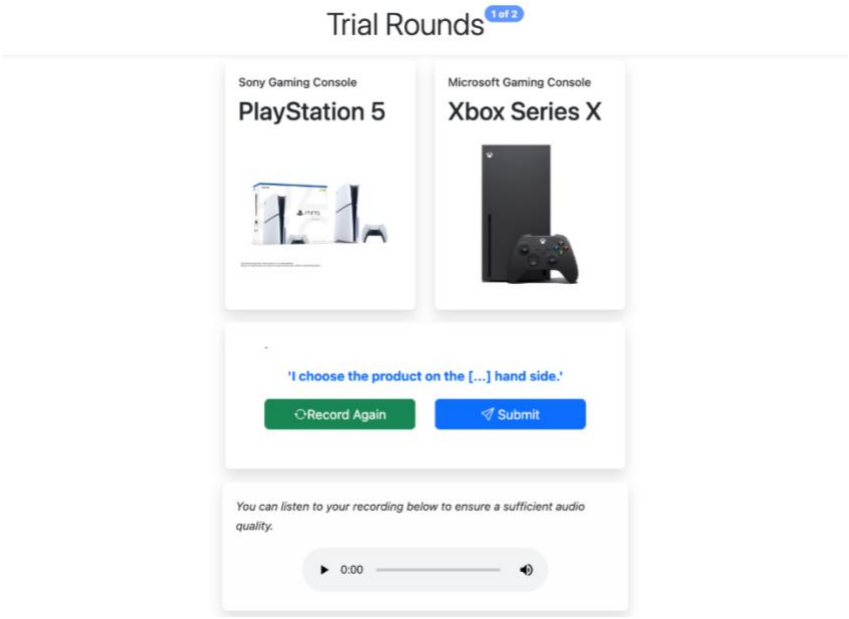
Trial Rounds (2 replicates, presentation order randomized)



Make sure that the audio is clear and understandable. You cannot be paid, if this is not the case. Furthermore, you will not be able to listen or re-take your recordings at later stages of this experiment.

[PAGE BREAK]

Audio Check



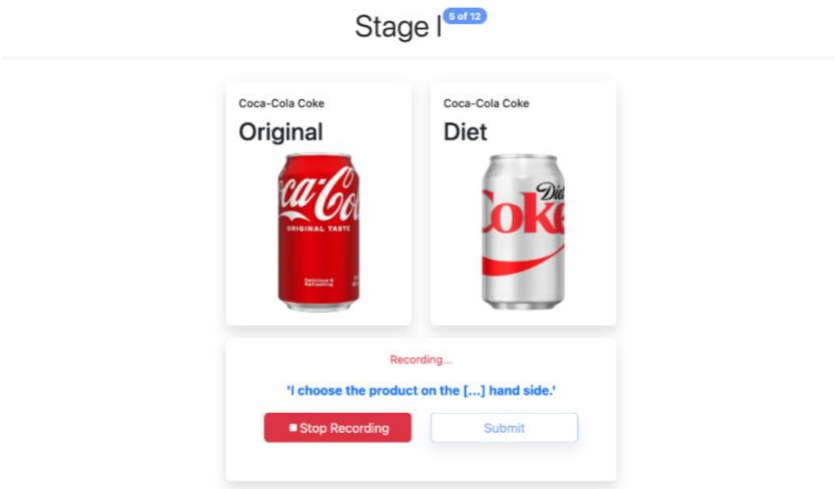
Stage I Instructions

The following pages will look similar: you will be presented with a series of product pairs and for each product pair, your task is to indicate the product you would prefer to choose by saying "I choose the product on the [left/right] hand side." aloud.

In this stage, however, you cannot re-take your recordings. In addition, the software will present a transcript after you stopped the recording. If applicable, you can indicate that the transcript does not reflect your choice. To proceed with the next product pair, you submit the page.

[PAGE BREAK]

<u>Choice Task</u> (12 replicates, order randomized)



[PAGE BREAK]

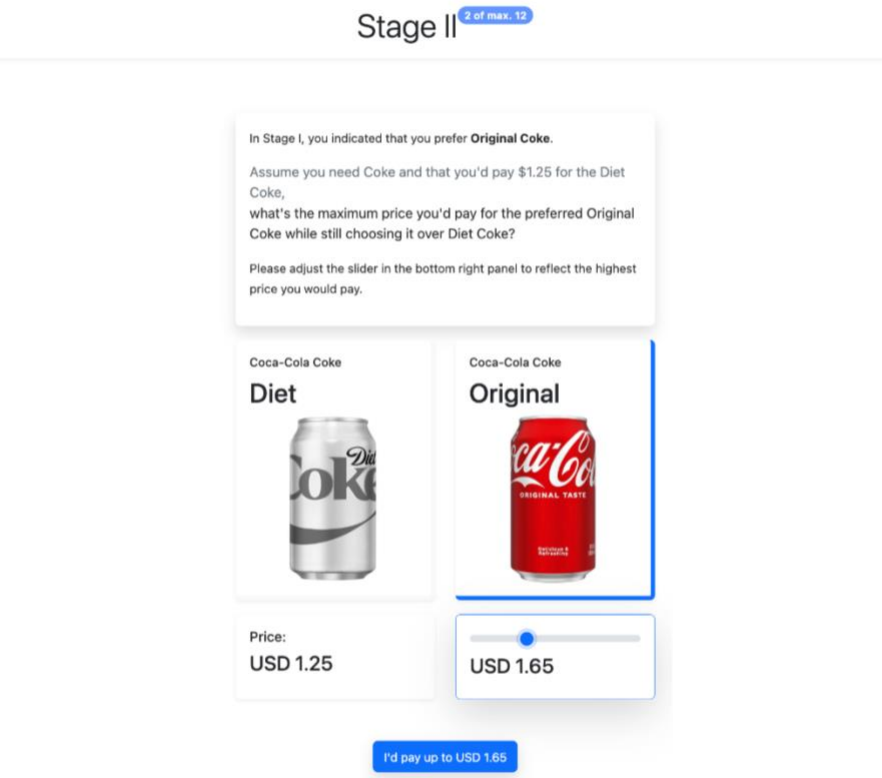
<u>Stage II Instructions</u>
On the following pages, you will see the same product pairs again. This time, we are interested in how much more you would be willing to pay for your preferred item.
Accordingly, we will ask you the following (exemplary) question for each product pair:

In Stage I, you indicated that you prefer Apples.
Assume you need Fruits and that you'd pay $0.95 for a Pear, what's the maximum price you'd pay for an Apple while still choosing it over a Pear?

By dragging a slider, your task will therefore be to indicate the maximum price you are willing to pay for your preferred item (as indicated in Stage I).

[PAGE BREAK]

<u>Price Premium Acceptance</u> (12 replicates, order randomized)



[PAGE BREAK]

Demographics

- Please enter your age:
- Please select your gender. (1 = Female, 2 = Male, 3 = Other, 4 = Prefer not to say)
- What was your total household income before taxes during the past 12 months? (1 = Less than $25,000, 2 = $25,000-$49,999, 3 = $50,000-$74,999, 4 = $75,000-$99,999, 5 = $100,000-$149,999, 6 = $150,000 or more, 7 = Prefer not to say)

[PAGE BREAK]

Please use the box below to share your thoughts about the study (e.g., any aspects you particularly liked or disliked) to help us improve it.

Web Appendix G. Pretest Results (Study 3)

We recruited 100 participants (M_{age} = 38.50, SD = 11.69; 47% female) from Prolific and employed a 18-replicates repeated measures design.

Participants were asked to indicate your preference between two products across a series of choices. They were provided with a 18 pairs of product pairs (see table below for full list). We counterbalanced presentation order to control for order effects. For each choice set, participants indicated the strength of their preference by drawing a 101-point slider (-50 = Strong preferencve for left product, +50 = Strong preference for right product). Lastly, participants reported age and gender (see Web Appendix F for survey instrument).

As displayed in the figure below, the preference strength ratings revealed large variety across the product pairs. We retained 11 product pairs for the main study (see Web Appendix H).

Table: Stimuli used in pretest

Coffee
Light Roast

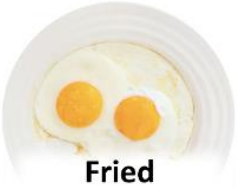
Coffee
Dark Roast

Splenda
Sweetener

Equal
Sweetener

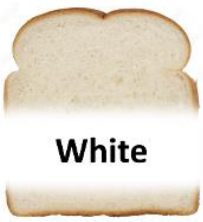
Coca Cola Diet

Coca Cola Zero



Fried
Eggs

Scrambled
Eggs



White

Whole Wheat

Still
Water

Sparkling
Water

Gum
Peppermint

Gum
Spearmint

Dish Soap
Orange

Dish Soap
Lemon

Orange Juice
No Pulp

Orange Juice
Some Pulp

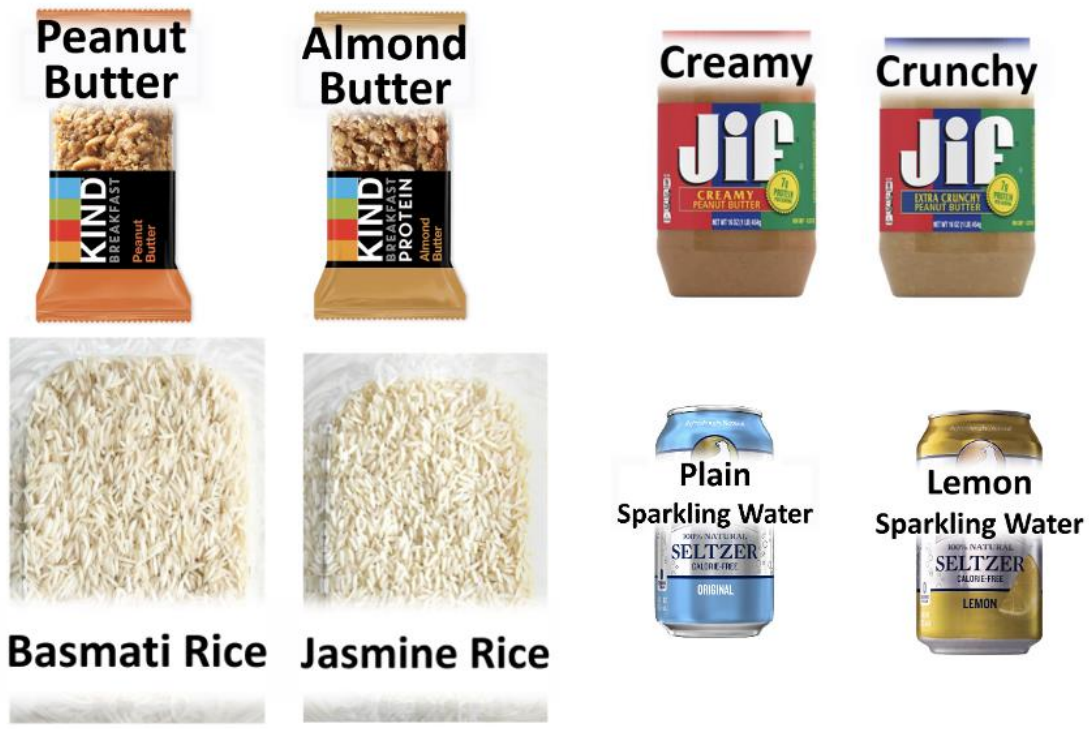
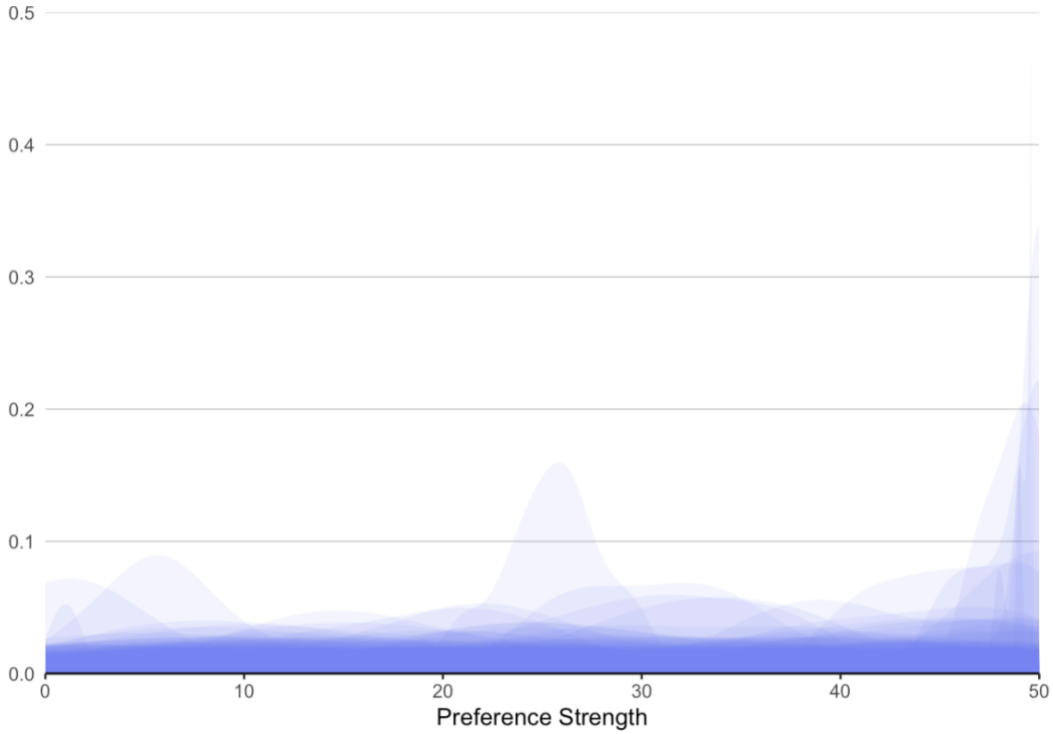


Figure: Distribution of Preference Strength per Participants Across Product Pairs



Web Appendix H. Survey Instrument (Study 4)

Consent

Dear Survey Participant:

This survey you are about to start should take you no more than 4 minutes to complete. You may only complete this survey once.

Please offer your candid opinions regarding the questions in this survey. We may have some questions to check that you were paying attention to the stimuli and to the other questions being asked.

There are no foreseeable risks associated with this project, nor are there any direct benefits to you. This information is anonymous and your identity will not be disclosed to anyone. The data will only be analyzed in aggregate. Your participation is voluntary, and you may withdraw from this project at any time. There is no penalty for doing so, but you will only receive payment if you complete the study.

If you have questions or comments, please contact us via email on prolific. If you want to talk to someone outside of the research team, you may contact the University of Pennsylvania Office of Regulatory Affairs (215-898-2614) with any questions or comments.

If you consent to participate in this study, please click "I agree to participate in this study".

[PAGE BREAK]

Instructions

On the following pages, we ask you to indicate your preference between two products across a series of choices.

You can indicate the strength of your preferences by moving a slider ranging from "Strong preference for left product" to "Strong preference for right product".

Below you can find examples that show different strengths of preferences:

1. Strong preference for left product

2. Weak preference for left product

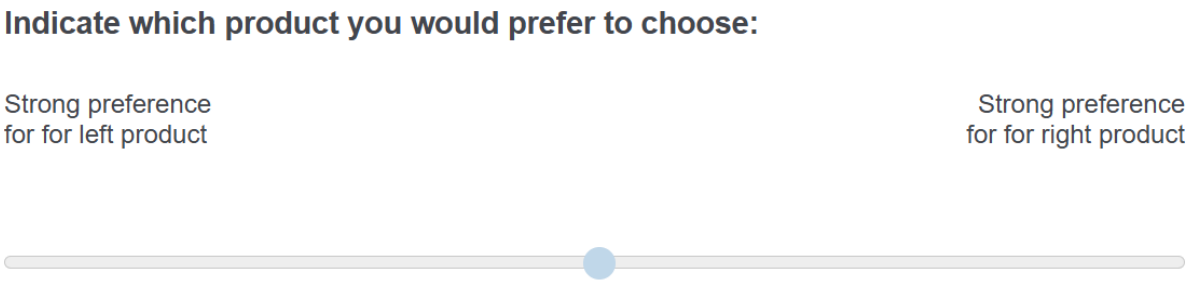
3. Weak preference for right product

[PAGE BREAK]

Preference Strength Task (18 replicates, presentation order randomized)

M&Ms
Milk Chocolate

M&Ms
Peanut

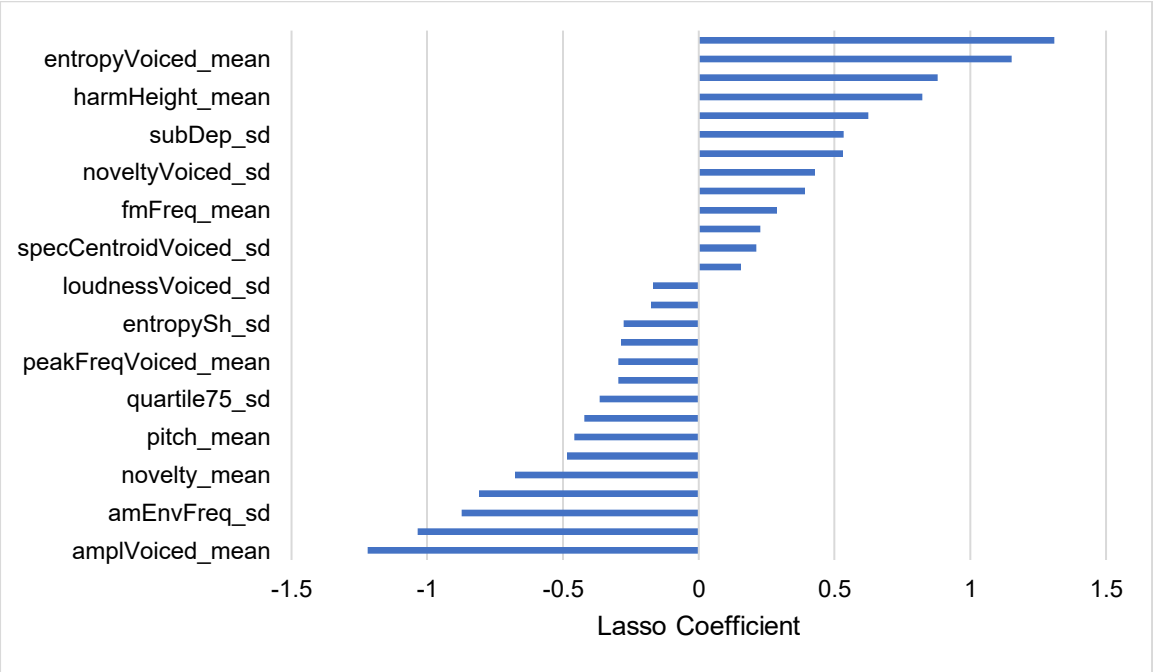


[PAGE BREAK]

Demographics

- What is your gender? (1 = male, 2 = female, 5 = other, 6 = prefer not to say)
- What is your age?

Web Appendix I. Results of LASSO Regression Predicting Anchored Price Premium (Study 3)



Web Appendix J. Survey Instrument Used in Pretest (Study 4)

Consent

Welcome!

Thank you for your interest in this study.

Study: Decision making study

Institution: XXXX

Duration: Approximately 6 minutes.

Payment: [Amount specified in session config]

Technical Requirements: We will ask you to provide a voice input. For that reason, we require you to use a device with a microphone in a quiet and private place. For the smoothest possible process, we advise using a modern browser other than Safari with a stable Internet connection.

Privacy:
☐ I agree to the anonymized analysis of my answers for the purpose of scientific evaluation according to the privacy policy.

[PAGE BREAK]

Instructions

Overview
This study is organized in two stages:

In Stage I, you will first have a short audio check that will familiarize you with the study setup.

Stage II is the main part of the study in which you have to make a series of short decisions: You will see several pages showing two options and you say out loud which of these two options you prefer.

Microphone Recording

In both stages, your microphone will automatically start recording at the start of every page. Your task is to first decide (i.e., say out loud) which option to choose and then click on the "Stop Recording"-button on the bottom of the page.

Procedure

In Stage I, there will be the option to listen to your voice recording again. You can re-record yourself in this first stage as often as you wish. This stage serves the purpose to make sure your audio is recorded properly. Please listen to your recording and ensure a high quality recording without background noise.

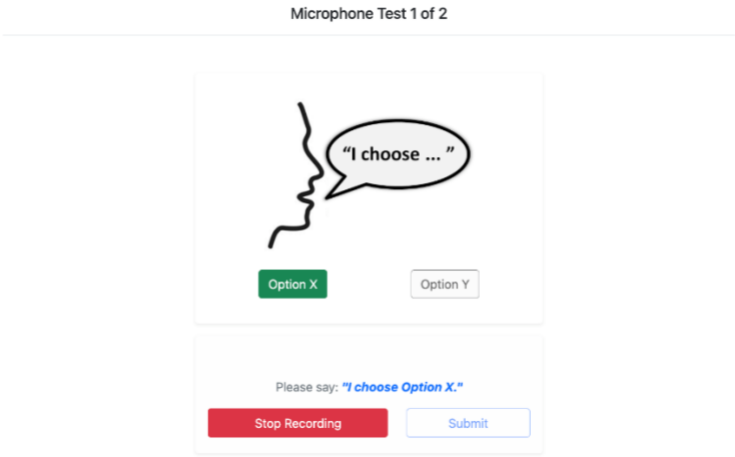
In Stage II, the recording starts automatically again (as in the first stage) but you will NOT have the option to listen to your recording again. In sum, please follow the following steps.

1. Express your preference by saying out loud 'I choose ...'.
2. Click on Stop Recording.
3. Only applies to Stage I: Check whether your recording was of high quality. If not, change your settings and try again.
4. Click on Submit to submit your voice input.

Once you have made all choices, we'll ask you to answer a short questionnaire to conclude this study.

[PAGE BREAK]

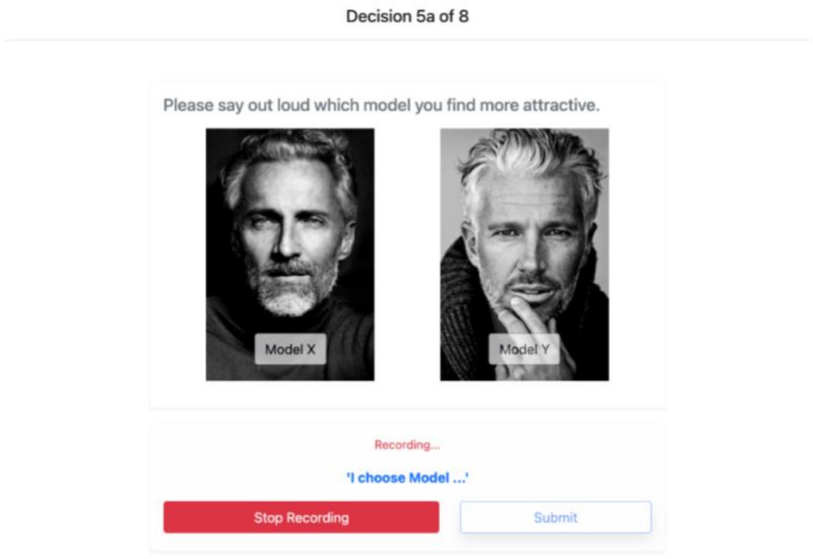
Microphone Test



Make sure that the audio is clear and understandable. You cannot be paid, if this is not the case. Furthermore, you will not be able to listen or re-take your recordings at later stages of this experiment.

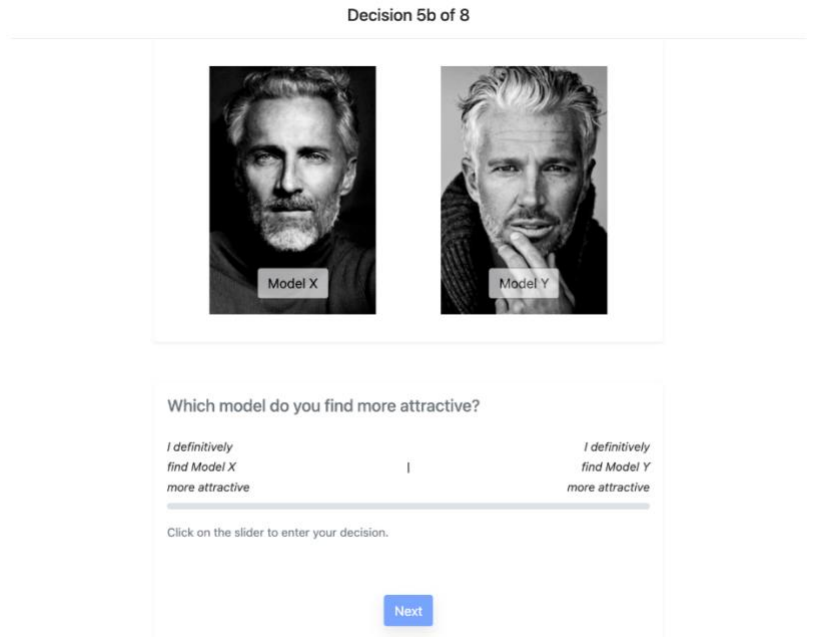
[PAGE BREAK]

Choice Task (8 replicates, presentation order randomized)



[PAGE BREAK]

Preference Strength



[PAGE BREAK]

Demographics

- Please enter your age
- Please select your gender. (1 = Female, 2 = Male, 3 = Other, 4 = Prefer not to say)

[PAGE BREAK]

Did you encounter any difficulties answering the study or do you have any comments?

Web Appendix K. Pretest Results of Female Stimuli (Study 4)

We recruited 131 postgraduate students (M_{age} = 24.28, SD = 2.76; 40% female) from a large European university and employed a 18-replicates repeated measures design.

Participants were asked to imagine that they are selecting models for an upcoming advertising campaign. They were provided with a 18 pairs of of female model headshots (see table below for full list). We counterbalanced presentation order to control for order effects. For each choice set, participants choose which woman would be perceived as more attractive by the majority of other people. Subsequently, for each pair, we measured decisional conflict by asking "How much conflict did you feel when choosing?" (1 = No conflict at all, 7 = A lot of conflict). Lastly, participants reported age and gender (see Web Appendix L for survey instrument).

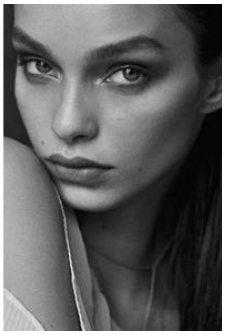
As displayed in the figure below, the conflict ratings revealed large variaty of decisional conflict across the headshot pairs. While some headhot pairs created only small decisional conflict, other headshot pairs produced high decisional conflict. We used the two headshot pairs that produced the highest (i.e., HC3 and HC9) and lowest decisional conflict (i.e., LC4 and LC6) for the main study.

Table: Stimuli used in pretest of female stimuli

HC1		
HC2		

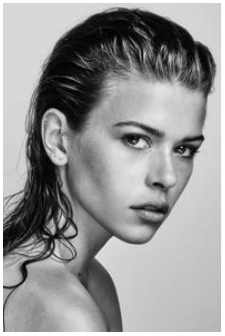
HC3


HC4

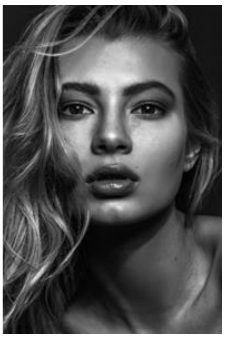
HC5


HC6

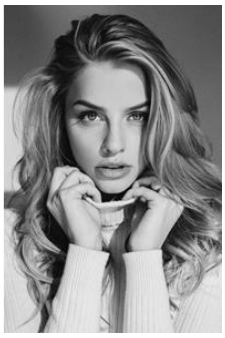



HC7


HC8

HC9

LC1


LC2


LC3


LC4


LC5

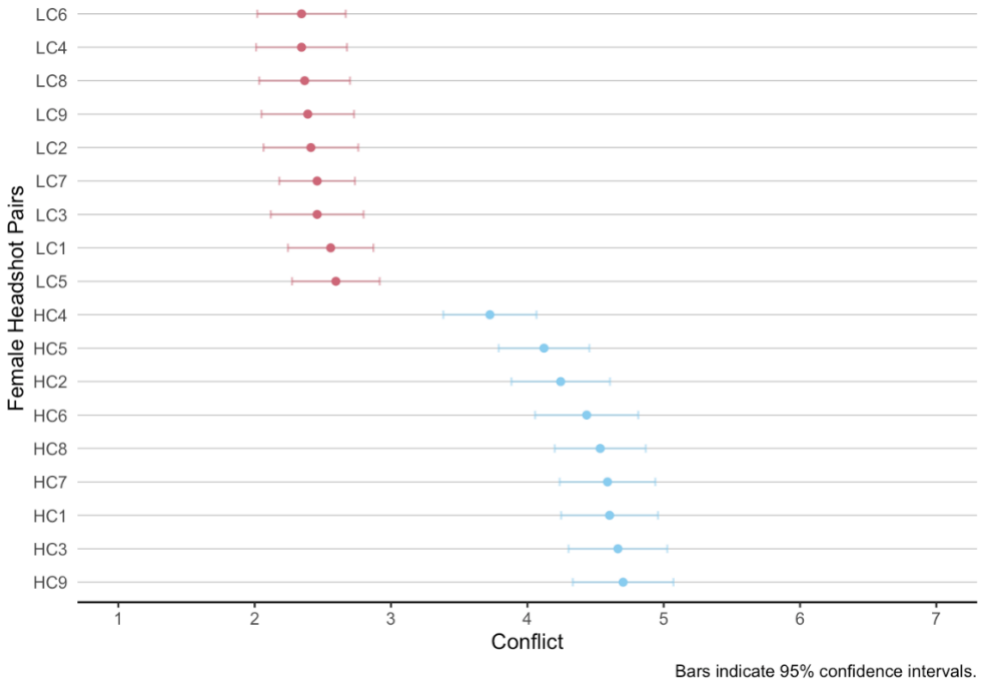

LC6


LC7


LC8


LC9

Figure: Decisional Conflict Results for Female Headshot Stimuli



Web Appendix L. Survey Instrument Used in Pretest of Female Stimuli (Study 4)

Consent

Dear Survey Participant:

The survey you are about to start should take you no more than 2 minutes to complete. You may only complete the survey once.

Please offer your candid opinions regarding the questions in the survey. We may have some questions to check that you were paying attention to the stimuli and to the other questions being asked.

There are no foreseeable risks associated with the project, nor are there any direct benefits to you. The information is anonymous and your identity will not be disclosed to anyone. The data will only be analyzed in aggregate. Your participation is voluntary, and you may withdraw from the project at any time. There is no penalty for doing so, but you will only receive payment if you complete the study.

If you have questions or comments, please contact us via email on prolific. If you want to talk to someone outside of the research team, you may contact XXXXXXXXXX with any questions or comments.

If you consent to participate in the study, please click "I agree to participate in the study".

[PAGE BREAK]

Intro

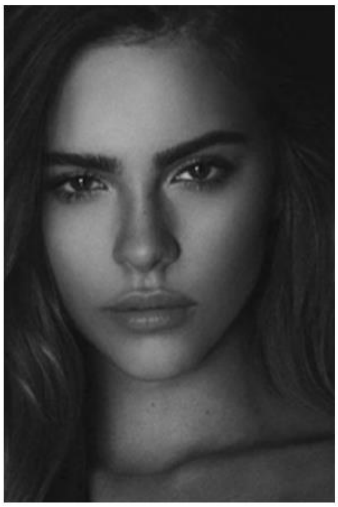
Imagine you are selecting models for an upcoming advertising campaign.

On the following pages, you will see a series of choices between two women. For each choice set, your task is to choose which woman would be perceived as more attractive by the majority of other people.

[PAGE BREAK]

Choice Task (18 replicates, presentation order randomized)

Which woman would the majority of people find more attractive? (example HC1)



Decisional Conflict

How much conflict did you feel when choosing? (1 = No conflict at all, 7 = A lot of conflict)

[PAGE BREAK]

Demographics

- What is your age?
- What is your gender?

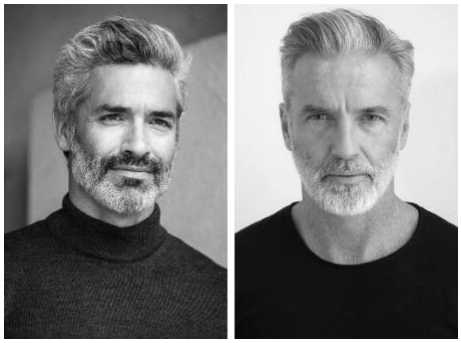
Web Appendix M. Pretest Results of Male Stimuli (Study 4)

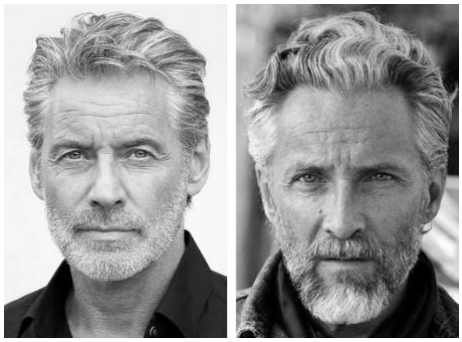
We recruited 100 participants (Mage = 39.72, SD = 14.24; 43% female) from Prolific and employed a 11-replicates repeated measures design.

Participants were asked to imagine that they are selecting models for an upcoming advertising campaign. They were provided with a 18 pairs of of female model headshots (see table below for full list). We counterbalanced presentation order to control for order effects. For each choice set, participants choose which woman would be perceived as more attractive by the majority of other people. Subsequently, for each pair, we measured decisional conflict by asking "How much conflict did you feel when choosing?" (1 = No conflict at all, 7 = A lot of conflict). Lastly, participants reported age and gender (see Web Appendix N for survey instrument).

As displayed in the figure below, the conflict ratings revealed large variaty of decisional conflict across the headshot pairs. While some headshot pairs created only small decisional conflict, other headhot pairs produced high decisional conflict. We used the two headshot pairs that produced the highest (i.e., HC4 and HC5) and lowest decisional conflict (i.e., LC4 and LC5) for the main study.

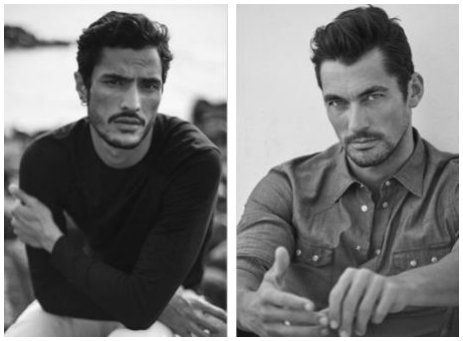
Table: Stimuli used in pretest of male stimuli

HC1 

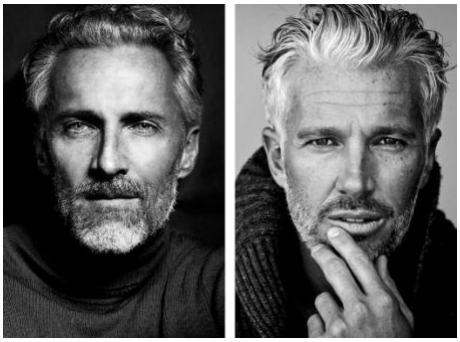
HC2 

HC3

HC4

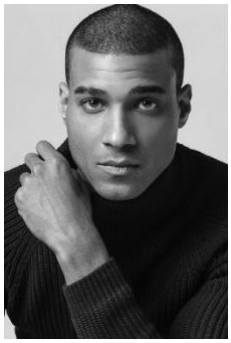


HC5



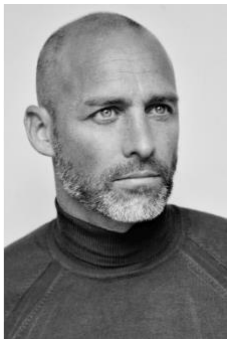
HC6

LC1




LC2

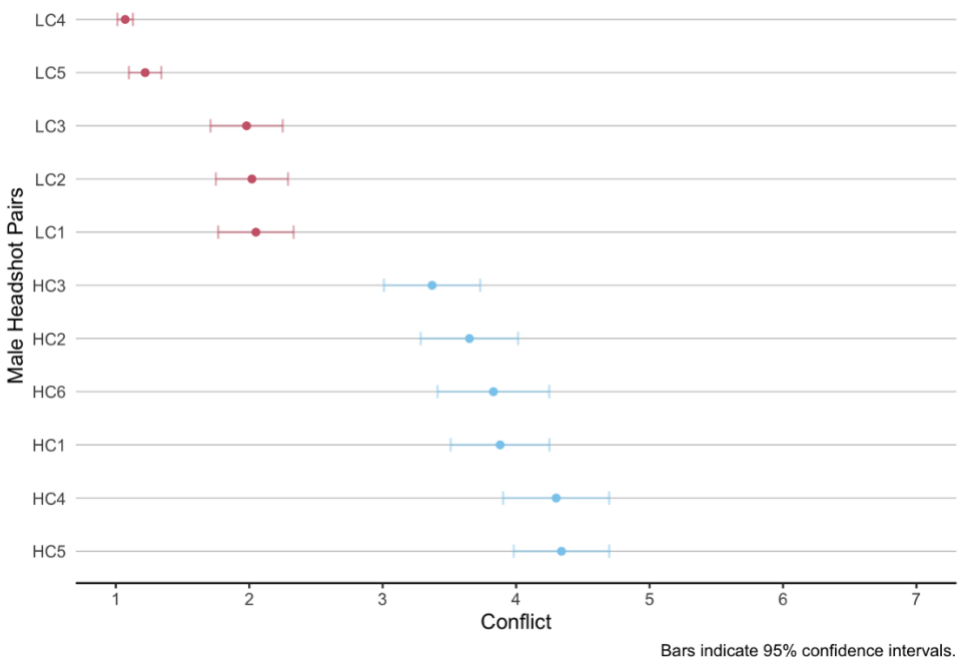

LC3

LC4


LC5


Figure: Decisional Conflict Results for Male Headshot Stimuli



Web Appendix N. Survey Instrument Used in Pretest of Male Stimuli

Consent Form

Dear Survey Participant:

The survey you are about to start should take you no more than 2 minutes to complete. You may only complete the survey once.

Please offer your candid opinions regarding the questions in the survey. We may have some questions to check that you were paying attention to the stimuli and to the other questions being asked.

There are no foreseeable risks associated with the project, nor are there any direct benefits to you. The information is anonymous and your identity will not be disclosed to anyone. The data will only be analyzed in aggregate. Your participation is voluntary, and you may withdraw from the project at any time. There is no penalty for doing so, but you will only receive payment if you complete the study.

If you have questions or comments, please contact us via email on prolific. If you want to talk to someone outside of the research team, you may contact XXXXXXXXXX with any questions or comments.

If you consent to participate in the study, please click "I agree to participate in the study".

[PAGE BREAK]

Intro

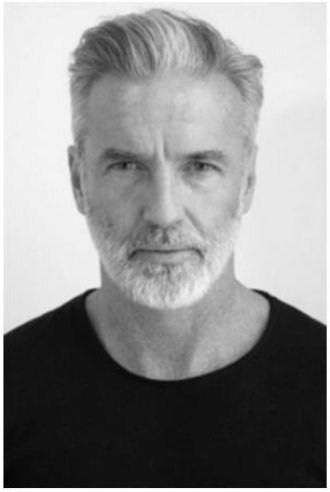
Imagine you are selecting models for an upcoming advertising campaign.

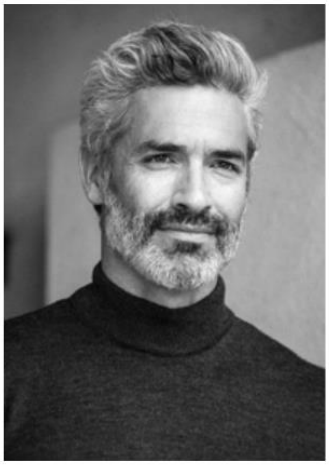
On the following pages, you will see a series of choices between two men. For each choice set, your task is to choose which man would be perceived as more attractive by the majority of other people.

[PAGE BREAK]

Choice Task (11 replicates, presentation order randomized)

Which man would the majority of people find more attractive? (example HC1)





Decisional Conflict

How much conflict did you feel when choosing? (1 = No conflict at all, 7 = A lot of conflict)

[PAGE BREAK]

Demographics

- What is your age?
- What is your gender?

Web Appendix O. Results of LASSO Regression Predicting Self-Reported Preference Strength (Study 4)