Check for updates

AMERICAN MARKETING ASSOCIATION

Article

Journal of Marketing Research
1-20
© American Marketing Association 2023
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/00222437221134115
journals.sagepub.com/home/mrj
SAGE

More Voices Persuade: The Attentional Benefits of Voice Numerosity

Hannah H. Chang, Anirban Mukherjee, and Amitava Chattopadhyay

Abstract

The authors posit that in an initial exposure to a broadcast video, hearing different voices narrate (in succession) a persuasive message encourages consumers' attention and processing of the message, thereby facilitating persuasion; this is referred to as the voice numerosity effect. Across four studies (plus validation and replication studies)—including two large-scale, real-world data sets (with more than 11,000 crowdfunding videos and over 3.6 million customer transactions, and more than 1,600 video ads) and two controlled experiments (with over 1,800 participants)—the results provide support for the hypothesized effect. The effect (1) has consequential, economic implications in a real-world marketplace, (2) is more pronounced when the message is easier to comprehend, (3) is more pronounced when consumers have the capacity to process the ad message, and (4) is mediated by the favorability of consumers' cognitive responses. The authors demonstrate the use of machine learning, text mining, and natural language processing to process and analyze unstructured (multimedia) data. Theoretical and marketing implications are discussed.

Keywords

voice, voice numerosity, videos, persuasion, marketing communications, sensory marketing, crowdfunding, advertising

Online supplement: https://doi.org/10.13039/501100001459

Video marketing often visually depicts a product, with one or multiple narrators discussing product features and benefits. For example, in Apple's video introducing its new AirPods Max, the voice-over has two narrators sequentially describing its features. Another video introducing the new MacBook Pro has voice-over by just one narrator. Upon consumers' initial exposure to such videos, does the number of voices narrating a message affect consumers' attention and processing of the message and subsequent behavior? If so, is the effect disruptive or facilitative? We examine these questions in the context of marketing communication videos (i.e., product videos and advertising), which have become increasingly prevalent and important in consumer decision making (Cramer-Flood 2021; Think with Google 2019).

Extant research from various disciplines suggests that sound—in particular the human voice—plays an important role in influencing consumer behavior (Horowitz 2012). Cognitive psychologists and neuroscientists argue that, compared with other sensory modalities, our brains have evolved to be exquisitely sensitive to the human voice (Belin, Fecteau, and Bédard 2004; Rutten et al. 2019). Babies can recognize it even though they do not yet understand language (Schweinberger et al. 2014). Hearing human voices activates distinct regions in the brain (Rutten et al. 2019; Von Kriegstein et al. 2010), quickly draws attention (Charest et al. 2009), and evokes immediate and greater processing (Aeschlimann et al. 2008).

Despite the importance of voice in cognition, existing research—spanning marketing, psychology, and information systems—has placed relatively little emphasis on understanding the influence of a narrator's voice in effective communications (Dahl 2010; Meyers-Levy, Bublitz, and Peracchio 2010). The limited work tends to study how specific audial features of a voice (e.g., volume) affect listeners' personal perceptions of the narrator, and thereby their attitudes and evaluations (e.g., Apple, Streeter, and Krauss 1979; Chattopadhyay et al. 2003; Wang et al. 2021). For example, a lower-pitched voice boosts persuasion through perception of increased competence

Hannah H. Chang is Associate Professor of Marketing, Lee Kong Chian School of Business, Singapore Management University, Singapore (email: hannahchang@smu.edu.sg). Anirban Mukherjee is Visiting Scholar, S.C. Johnson Graduate School of Management, Cornell University, USA (email: am253@cornell.edu). Amitava Chattopadhyay is GlaxoSmithKline Chaired Professor of Corporate Innovation and Professor of Marketing, INSEAD, France (email: amitava.chattopadhyay@insead.edu).

(Chattopadhyay et al. 2003; Wang et al. 2021). The findings highlight yet another challenge in the design of voice in marketing communications: while features of a narrator voice can increase persuasion by drawing consumers' attention to narrator cues, focusing on the narrator reduces attention to the spoken product message (Chaiken and Eagly 1983; Grewal, Gupta, and Hamilton 2021).

Moreover, prior research mostly focuses on the consequence of voice processing when only audial (vocal) information is provided (e.g., Craik and Kirsner 1974; Moore, Hausknecht, and Thamodaran 1986). Thus, even less is known about the persuasive effect of narrator voices when both visual and audial information are presented simultaneously, such as in product videos and video advertising. This is important to understand as consumers increasingly embrace video content in making consumption decisions (Cramer-Flood 2021; Think with Google 2019). For example, more than 90% of consumers rely on online product videos to discover new brands or products, and more than 50% use videos to decide which specific brand or product to buy (Think with Google 2019). Such videos typically employ narrator voices to convey focal messages. In particular, a widely used production technique in video marketing is voice-over narration, whereby information about a brand or product is spoken by one or more off-screen narrators. It is found in 89% of TV ads (Millward Brown 2012) and 80% of product videos posted by brands on social media platforms (Facebook IQ 2019).

In this research, we hypothesize that in an initial exposure to a video ad, consumers who hear a persuasive message narrated by different voices (in succession) are likely to be more persuaded than consumers who hear the exact same message narrated by one voice.[1] We attribute this effect to consumers' natural predisposition to attend to the human voice (Belin, Fecteau, and Bédard 2004; Charest et al. 2009) and to the fact that a change in voice can involuntarily capture attention, even when other visual or auditory tasks are competing for attention (Cherry 1953; Morton, Crowder, and Prussin 1971). Therefore, in a video with multiple narrating voices where a new narrator's voice carries on the persuasive message, the change in voices should help facilitate processing of the spoken message, boosting persuasion. We term this phenomenon the voice numerosity effect.

Consistent with this proposition, we report four studies (plus validation and replication studies), including two large field studies leveraging data from a leading crowdfunding platform (Study 1, with more than 11,000 product videos and over 3.6 million customer transactions) and online video advertisements (Study 2, with more than 1,600 ads) and two experiments (Studies 3 and 4, with over 1,800 participants). The results show that the voice numerosity effect is robust. It is observed across different contexts (product categories, advertising topics), types of video marketing (product videos, advertisements), and narrator voices (humans, machine-synthesized). It can improve a wide range of behavioral outcomes relevant to marketing practice, including consequential crowdfunding project outcomes, perceived advertising efficacy, consumers' willingness to pay (WTP) for a target product, and purchase likelihood. Moreover, the findings circumscribe the conditions under which voice numerosity is more likely to facilitate persuasion. By examining different moderators of voice numerosity and process mediation, we provide converging evidence that the effect is due to increased consumer attention and processing.

We contribute to the marketing literature on the effective design of voice in video marketing, in terms of both theory and practice. We add to the literature by showing that the number of narrating voices in videos—even in the presence of other visual and audio information—can affect consumers' information processing and behavior. We also demonstrate the use of machine learning and natural language processing (NLP) to overcome methodological challenges previously identified as reasons for the paucity of research on the impact of voice on consumer behavior (see Grewal 2018; Krishna and Schwarz 2014) despite the extensive use of voices in marketing. To gain insights into current practice, we interviewed five senior executives (two from global ad agencies and three from global fast-moving consumer goods businesses), who revealed that video narration with a single voice or with multiple voices is common in practice but employed in a nonstrategic manner; the executives were not aware of our hypothesized effect prior to the interviews. Therefore, our findings offer guidance to practitioners on the design of video narration.

Theoretical Background

Prevalence of Videos in Marketing Communication

Marketers use a variety of media (e.g., written text, audio, image, video) to communicate to consumers, whether to persuade individuals to buy a new product, watch a newly released movie, vote for a presidential candidate or the next "idol," or support a social cause. Marketing media consist primarily of two types of information: visual and audio. Some communications contain only one type of information, such as print ads (visual) and radio ads (audio); others include both types, such as television ads (audiovisual). In recent years, the proliferation of digital technology has dramatically increased the prevalence and variety of video marketing beyond television. For example, brands like Apple, BMW, and Lego post videos about their products on their respective YouTube channels and social media accounts. Meanwhile, consumers are increasingly embracing video consumption; the number of digital video viewers worldwide reached 3 billion in 2020 and was projected to reach nearly 3.5 billion by 2023 (Statista 2023). Various digital forms of product videos and broadcast ads are gaining prominence in consumer decision making (Think with Google 2019).

[1] In this research, we consider any voice (off-screen or on-screen narrator) that appears in a video's audio track.

Despite the growing popularity of audiovisual content, existing research—spanning marketing, psychology, communications, and information systems—has placed less emphasis on the design of narrator voice (audio) in communications (Dahl 2010; Meyers-Levy, Bublitz, and Peracchio 2010). This oversight is perhaps due to the methodological challenges entailed in (1) designing appropriate stimuli (see Krishna and Schwarz 2014) and (2) extracting and analyzing audiovisual data (see Grewal 2018). We leverage advances in machine learning techniques to overcome these challenges (described in our studies). We next turn to a review of prior research on the acoustic elements of marketing media.

Acoustic Elements in (Asynchronous) Broadcast Videos

Acoustic elements in asynchronous broadcast videos, such as product videos and TV ads, typically consist of (1) background music (Edell and Burke 1987; Zhu and Meyers-Levy 2005) and (2) the narrator's voice (Chattopadhyay et al. 2003; Forehand and Perkins 2005). Research shows that these elements can significantly influence product perception (Zhu and Meyers-Levy 2005), purchase intention (Alpert and Alpert 1990), brand evaluation (Anand and Sternthal 1990), brand beliefs (Edell and Burke 1987), and ad attitude (Chattopadhyay et al. 2003).

Background music is the more commonly investigated acoustic element of broadcast videos (Alpert and Alpert 1990; Anand and Sternthal 1990; Edell and Burke 1987; Zhu and Meyers-Levy 2005). Central to our research is the second type of acoustic element: the narrator's voice. The importance of voice in conveying ad messages has been widely acknowledged in practitioners' guides to video marketing. Facebook IQ (2019) recommends to businesses that video ads be enhanced with voice-over to increase ad effectiveness. YouTube Advertising (2019) suggests that the focal product be introduced with "a clear speaking voice" in making a video ad. A 2018 industry report indicates that in creating ad campaigns and videos, marketing and advertising professionals believe that voices signaling authority and relatability best resonate with consumers (Voices 2018). Yet, practical considerations for choosing a voice for the ads are driven mostly by practitioners' intuition (see also Chattopadhyay et al. 2003).

There is limited research to provide theory-driven guidance on the "voice" element in effective communications (Dahl 2010; Meyers-Levy, Bublitz, and Peracchio 2010). As Dahl (2010) points out, "Despite the apparently important role of voice in determining the effectiveness of a broadcast advertisement, little research has been done in this area" (p. 170). Extant research focuses on how the specific qualities of a voice may impact consumer behavior. For example, Forehand and Perkins (2005) find that when consumers recognized (did not recognize) the celebrity voice-over, their explicit brand attitudes were negatively (positively) related to their attitude toward the celebrity. Chattopadhyay et al. (2003) find that a voice with lower pitch, and a voice with faster syllable speed, led to more favorable brand attitude in radio ads, whereas interphrase pausation had little effect. Wang et al. (2021) find that a persuader's voice (signaling focus, low stress, and stable emotions) can exert influence through person perception of competence (see also Brown, Strong, and Rencher 1974). These results indicate that features of a voice in broadcast ads can systematically affect consumer information processing and attitude. In this article, we build on prior research to investigate whether and how voice features beyond those specific to a narrator's voice can impact consumer behavior, in the presence of other audiovisual content (i.e., videos).

The Voice Numerosity Effect

We propose that in an initial exposure to a marketing communication video, hearing different voices (vs. the same voice) narrate a persuasive message can affect consumers' attention and processing of the message, and thereby its persuasive appeal. Research in cognitive psychology and neuroscience suggests that, across sensory modalities, humans are sensitive to sound (Horowitz 2012) and to the human voice in particular (Rutten et al. 2019), among the myriad sensory signals in our environment. An emerging body of neuroscience studies on multisensory (audiovisual) content shows (1) the dominance of auditory processing (Robinson, Moore, and Crook 2018; Robinson and Sloutsky 2019) and (2) the superior processing speed of hearing, compared with processing through other senses such as vision (Horowitz 2012).

The human voice is the most important sound in our environment (Belin, Fecteau, and Bédard 2004; Grossmann et al. 2010); it not only conveys speech (Rutten et al. 2019) but carries socially relevant information for communication (Belin, Fecteau, and Bédard 2004). Hearing human voices activates distinct regions in the brain (Rutten et al. 2019), quickly draws attention due to stimulus significance (Charest et al. 2009), and evokes immediate and more processing (Aeschlimann et al. 2008). Even the brains of infants show heightened sensitivity to the human voice (Grossmann et al. 2010). These converging findings have led neuroscientists to suggest a long evolutionary history underlying voice-preferential processing (Charest et al. 2009; Petkov et al. 2008).

Consumers cannot attend to all incoming sensory stimuli (Nosofsky 1984), and thus they consider a subset of features that capture their attention (Hoffman and Singh 1997). For example, visual salience of a part (in figure or ground) can capture attention and influence what people see in a figure–ground image (Hoffman and Singh 1997). Visual selective attention can also drive subsequent choices (Janiszewski, Kuo, and Tavassoli 2013). A key aspect of a stimulus that captures consumer attention is stimulus change. For example, Furedy and Scull (1971) found that participants who are exposed to an unpredictable sequence of events (shock and cool-air puff) displayed a greater orienting response with a change in event than with repetition of the same event (see also Gati and Ben-Shakhar 1990; Kahneman 1973; Sokolov 1963). This is because preattentive mechanisms detect the change and issue a "call" for reallocation of central processing resources (Öhman 1979). Thus, we posit that a voice change in narration should capture consumer attention.

Consistent with our theorizing, prior studies demonstrate that a change in voice can involuntarily capture attention, even when there are other visual or audial tasks competing for attention (Cherry 1953; Morton, Crowder, and Prussin 1971; Treisman and Riley 1969). For example, in a seminal study by Cherry (1953), participants who listened to two different messages concurrently were able to shadow the focal message (i.e., repeat the heard message aloud) and ignore the nonfocal message. When there was a change from one person's voice to another in relaying the nonfocal message, participants easily and consistently noticed the change in voice, demonstrating (at least) a temporary shift in attention due to the change in voice. Research in neuroscience on the impact of stimulus change in auditory information on attention, using event-related potential measures, has shown that sounds that are novel or unexpected can involuntarily capture attention, and do so by activating different neural mechanisms in the brain (Escera et al. 1998). Other neuroscience studies have shown that certain brain regions (e.g., in the anterior temporal lobe) respond more vigorously to speech from different speakers than to speech from the same speaker (Belin and Zatorre 2003; Von Kriegstein et al. 2010).

Taking these findings together, we hypothesize that in an initial exposure to a broadcast video, hearing different voices narrate (in succession) a persuasive message can foster persuasion, compared with hearing the same voice narrate the exact same message. In a video wherein the spoken narration has multiple voices and a new narrator's voice carries on the persuasive message, the change in narrator voice should promote consumers' (continued) attention and processing of the next piece of spoken message, which might not have been processed otherwise. The enhanced attention and processing of the persuasive message would facilitate its persuasive appeal. We term this phenomenon the voice numerosity effect. From a message-processing point of view, once the recipient's attention is obtained, the recipient needs to allocate processing resources for the attentional advantage to translate into greater persuasive impact. Drawing on well-established frameworks, we posit that the effect is more likely when consumers have greater opportunity and ability to process the message (MacInnis and Jaworski 1989; Petty and Cacioppo 1986).

Overview of Studies

We test our predictions in four studies—with two large real-world data sets and two controlled experiments (plus validation and replication studies)—spanning diverse decision domains, product categories, voice-based marketing communication tools, and a wide range of outcomes. The studies also examine our conceptualization by testing the boundaries of the hypothesized effect under two theoretically derived moderators: opportunity (Studies 1 and 2) and ability (Study 3) to process information (see MacInnis and Jaworski 1989). We find that the effect is mediated by the favorability of cognitive responses toward the product (Study 4).

Study 1 examines the hypothesized effect using a comprehensive data set we collected on Kickstarter, a leading crowdfunding platform where product videos are commonly used to communicate new products to potential consumers (Mollick 2014). We apply machine learning, NLP, and text mining to process the unstructured multimedia data for our investigation of voice numerosity on consequential outcomes in crowdfunding. Study 2 extends our test of the effect to video ads, which is an important marketing communication tool; spending on video ads in the United States is expected to reach U.S. $145 billion by 2023 (Perrin 2021). We augment a data set of video ads obtained from Hussain et al. (2017) to test the effect on perceived efficacy of ads. Previous studies have shown that faster speech rates disrupt listeners' cognitive processing (Goldinger, Pisoni, and Logan 1991; Moore, Hausknecht, and Thamodaran 1986; Smith and Shaffer 1995). Thus, in Studies 1 and 2, we assess whether the effect is moderated by speech rate (i.e., processing opportunity); to mitigate endogeneity concerns associated with real-world data sets, we report an extensive set of 24 robustness checks and utilize propensity scores to create balanced samples (Rosenbaum 2020). Studies 3 and 4 are controlled experiments to further improve causal inference. Study 3 tests the voice numerosity effect under varied distractions (which affect processing ability). Study 4 measures consumers' cognitive responses to examine the underlying process.

Study 1: Voice Numerosity in Crowdfunding

We situated our first empirical investigation of the voice numerosity effect in online crowdfunding, which is an important avenue to drive the adoption of new products (Dhanani and Mukherjee 2017). We collected data from Kickstarter, a leading crowdfunding platform, on which entrepreneurs and companies launch "projects" that are described on web pages to presell new products. Potential customers (or "backers") browse the project web page to determine if they will prepurchase the product and support the project. The platform provides a suitable empirical context to test our hypothesized effect. First, a project video is among the first things potential customers see on a project web page, and it is vital for customer conversion (Pollari 2015). Second, prior studies note that a majority of the projects on Kickstarter (about 80% to 86%) include product videos (Gafni, Marom, and Sade 2020; Mollick 2014). Third, the platform is a self-contained marketplace, with not only a comprehensive description of marketing communication messages directed at consumers but also a detailed account of consequential consumer behavior.

The purpose of this study was twofold. First, we aimed to examine the voice numerosity effect in a marketplace with real-world consumer behavior and consequential dependent variables. Second, we aimed to test our conceptualization that the hypothesized effect relates to consumers' cognitive processing. Building on prior studies showing that a faster speech rate

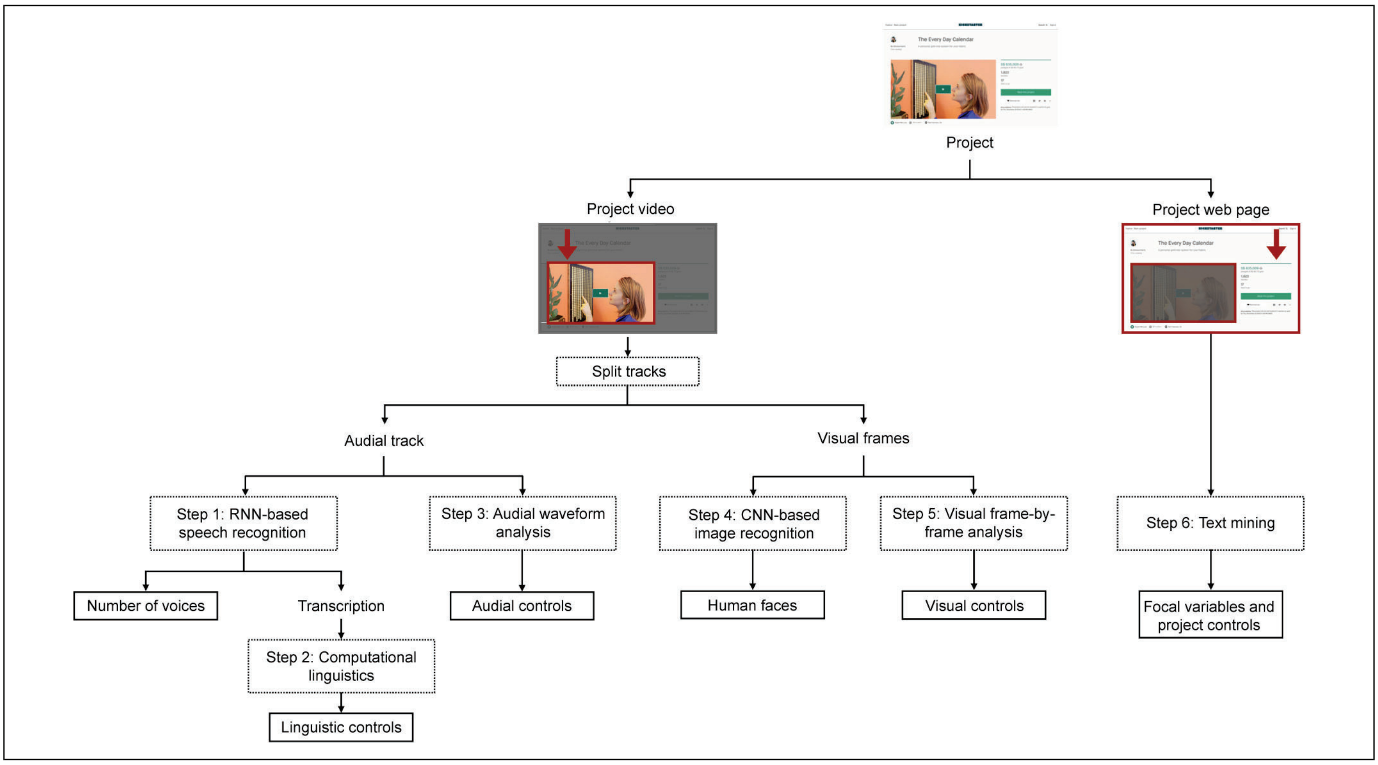
Project
Project video
Project web page
Split tracks
Audial track
Visual frames
Step 1: RNN-based speech recognition
Step 3: Audial waveform analysis
Step 4: CNN-based image recognition
Step 5: Visual frame-by-frame analysis
Step 6: Text mining
Number of voices
Transcription
Audial controls
Human faces
Visual controls
Focal variables and project controls
Step 2: Computational linguistics
Linguistic controls

Figure 1. Overview of Variable Construction for Crowdfunding Data.

disrupts listeners' cognitive processing (Goldinger, Pisoni, and Logan 1991; Moore, Hausknecht, and Thamodaran 1986), we assessed whether the voice numerosity effect is moderated by speech rate. We predicted that having more voices narrate a product message enhances its persuasiveness and thereby improves project outcomes when the message is spoken at a slower rate (when speech is easier to process) but not when the message is spoken at a faster rate (when speech is more difficult to process).

Research Setting and Data

We obtained a preliminary data set from WebRobots (https://webrobots.io/kickstarter-datasets/) that includes the project URL and some basic information on each project (e.g., project name, country) for all projects on Kickstarter from July 1, 2017, to December 31, 2019, in 31 categories corresponding to the three largest supracategories on Kickstarter: Design, Games, and Technology (see Table W1 in Web Appendix A). These supracategories account for almost 70% of the funding raised on Kickstarter (Statista 2021). The WebRobots data set does not include project web pages and product videos, which we needed to construct focal variables based on our conceptual framework. We collected these directly from Kickstarter. Our sample has 11,801 U.S.-based projects that included a project video, with a total of over 3.6 million customer transactions and more than U.S. $382 million in pledged funding. Table W2 in Web Appendix A presents the summary statistics.

Parsing Procedure, Operationalization, and Measures

The raw data of the crowdfunding projects (videos and web pages) are unstructured. Using human coders to derive relevant focal and control variables for 11,801 videos and project web pages would be difficult and cost-prohibitive. To overcome this challenge, we used cutting-edge methods from machine learning and computer science to algorithmically code the variables.

To prepare the raw video data for processing, for each video we decoupled the audial track and visual frames. In Step 1, we processed the audial tracks to (1) identify the number of voices narrating the message and (2) obtain a text transcription of the spoken words. In Step 2, we constructed psychologically relevant measures of the linguistic tone of the speech transcription. In the subsequent steps (3 to 6), we measured audial, visual, and project controls following extant literature in marketing, psychology, crowdfunding, and information systems. In Step 3, we measured variables to account for differences in audial characteristics, as suggested by prior literature (e.g., Li, Shi, and Wang 2019; Packwood 1974). In Step 4, we focused on the video track and used a face-detection model to measure the prevalence of human faces in the visual frames to detect possibility of on-screen communicators. In Step 5, we developed measures that relate to the frame-by-frame, pixel-by-pixel variation to account for visual characteristics, following recent studies (e.g., Liu et al. 2018). In Step 6, we analyzed the text descriptions of the project pages to derive project controls from the project web pages (e.g., Li, Shi, and Wang 2019). Figure 1 is a graphical overview of our parsing procedure and variable construction. We next detail these steps and the variables.

Step 1: Speech recognition for number of voices. We applied an automatic speech-recognition model (ASR; Makino et al. 2019) to the audial data. It operates directly on the waveform of the audio track to detect "who said what" through speaker diarization. ASRs (1) detect which part of the acoustic waveform relates to speech; (2) detect when speakers in the conversation change through shifts in the acoustic characteristics of speech (e.g., timbre); and, (3) through the acoustic signature, identify what was said by each speaker in the entire conversation.

Modern deep-learning-based ASRs, such as the state-of-the-art ASR from Google that we utilized, employ recurrent neural networks (RNNs)—a class of machine learning models—to accomplish these tasks (Graves, Mohamed, and Hinton 2013). RNNs use large latent state spaces to capture long-term dependencies in sequential data. Speech is sequential in nature, as the current spoken word depends on both what was spoken before it and what will be spoken after it. RNNs infer the word that was spoken from both the sequence of spoken syllables and the sequence of spoken words. Thus, RNN-based ASRs are able to achieve greater computational efficiency and accuracy than conventional ASRs (Chelba et al. 2013). RNN-based ASRs were developed and refined through repeated verifications with human judgments for enhanced accuracy (Flaks et al. 2018). To ensure that the machine-coded number of voices in the audio track aligns with human speech perception, we conducted a validation study (which we discuss subsequently). The results validate the machine-coded measure of number of voices.

The ASR enables us to derive our focal independent variable—the number of voices in each audio track—across 11,801 videos in our sample. The raw data, however, include voices that spoke few words (e.g., "wow"), which a typical human audience would likely not consider a "speaker" but background sound. As prior findings on the average sentence length in spoken English range from 12.9 words (Vajjala and Meurers 2014) to 16.6 words (Poole and Field 1976) to 17.9 words (O'Donnell 1974), we include voices that spoke at least 13 words in our machine-coded measure.[2]

Step 2: Computational linguistics for linguistic controls. We employed the ASR (Step 1) to derive a text-based transcription of spoken content for each product video. A recent and growing literature in marketing investigates the linguistic characteristics of verbal text in relation to consumer behavior (e.g., Cavanaugh, Bettman, and Luce 2015; Melumad, Inman, and Pham 2019). Following prior studies, we measured the linguistic tones in the spoken content to the extent that they have been previously linked to persuasion.

To prepare the raw speech transcriptions for data analysis, in line with common practice, we preprocessed the transcriptions as follows. We converted text to ASCII to remove special characters, converted text to lowercase, replaced contractions (e.g., "don't" becomes "do not"), and removed all punctuation marks. Finally, to ensure that we included only English words and names, we retained words that appear in either of two state-of-the-art lists: Grady Ward's list of English words and Mark Kantowitz's list of English names.[3]

We applied Linguistic Inquiry and Word Count (LIWC; Pennebaker et al. 2015) to the preprocessed data. LIWC is a seminal text-analysis method based on statistical models that link the use of words and phrases in verbal communication to higher-order psychological constructs such as agreeableness in the message sender and recipient; it has been used extensively in social psychology and consumer behavior research to measure psychological constructs from textual data, such as online blogs, customer reviews, and text messages (e.g., Cavanaugh, Bettman, and Luce 2015).

To account for the linguistic characteristics of the spoken message, we utilize four summary language variables from LIWC2015: (1) analytical thinking (which captures logical thinking patterns in verbal content), (2) authenticity (which measures the extent to which verbal content displays honest, authentic discourse), (3) clout (which identifies speech related to high expertise and confidence), and (4) emotional tone (which gauges affective tones in language use) (Pennebaker et al. 2015). Communications exhibiting these linguistic characteristics have been previously linked to persuasion. For example, positive mood and emotional ads can increase purchase intentions (e.g., Alpert and Alpert 1990), and authenticity can increase customers' WTP (e.g., Lehman, O'Connor, and Carroll 2019). We also included the number of words as is common in computational linguistics (e.g., Melumad, Inman, and Pham 2019), as more information has been linked to persuasion (Calder, Insko, and Yandell 1974).

Step 3: Waveform analysis for audial controls. From the waveforms of the audio tracks, we measured audial characteristics that researchers have linked to persuasion (Packwood 1974) or used as audial controls (Li, Shi, and Wang 2019). Prior studies have shown that louder recording (Oksenberg, Coleman, and Cannell 1986; Packwood 1974) and longer speech (Moore, Hausknecht, and Thamodaran 1986) boost persuasion. Central to our theorizing, prior work has shown that faster speech rate disrupts listeners' cognitive processing (Goldinger, Pisoni, and Logan 1991; Moore, Hausknecht, and

[2] We repeated the analyses with (1) the other two estimates of average sentence length in spoken English from prior studies, (2) different word counts for cutoff (e.g., one word, two words), and (3) no cutoff (i.e., a voice saying only "wow" as a unique speaker). Our results are robust across these thresholds in defining a voice for human speech processing.

[3] Grady Ward's list of English words can be found at http://www.gutenberg.org/files/3202/files, and Mark Kantowitz's list of English names is at http://www.cs.cmu.edu/afs/cs/project/ai-repository/ai/areas/nlp/corpora/names/.

Thamodaran 1986; Smith and Shaffer 1995). We thus measured the average volume (in decibels) and the duration (in seconds) of each audio track, from which we derived the rate of speech (in words per second) (Mairesse et al. 2007). We also computed audial controls following Li, Shi, and Wang (2019), including audial entropy (change in the energy of the audio), energy (short-term dynamism of the waveform), spectral centroid (brightness of the sound), spectral entropy (change in the brightness of the sound), and zero crossings (the noisiness of an audio signal).

Step 4: Face detection for visual controls. To account for the possible visual presence of communicators, we applied a convolutional neural network (CNN) face-detection model by Google to visual frames. Each project video has 24 frames per second. As each frame is an image, processing the entire video is prohibitively costly (if coded by humans) and computationally intractable[4] (if coded by algorithms, due to the large size of the images). Fortunately, analyzing every frame is unnecessary, as neighboring frames convey similar information (Smith and Kanade 1998). Recent research in marketing tends to sample, code, and analyze (1) only the first frame of each video (Ordenes et al. 2019) or (2) three frames, from the beginning, middle, and end (Li, Shi, and Wang 2019). To enhance precision of visual controls given the computational constraints, we sampled ten frames from each video, one at the median of each decile of the visual frames. We applied the face-detection model (a state-of-the-art CNN model trained over billions of training samples) to identify the presence of human faces in the sampled frames. The model has been repeatedly validated. For example, Li and Xie (2020) found that it achieves an accuracy rate of 95% and a precision of 92.7% in human face detection, relative to human coders.

Step 5: Image analyses for visual controls. We analyzed the frame-by-frame, pixel-by-pixel characteristics of all visual frames to account for visual features that may affect consumer attention. Specifically, number of scenes is a discrete measure to characterize the amount of visual information, based on a conceptualization of scenes as building blocks of videos (see Liu et al. 2018). Visual variation is a measure by Li, Shi, and Wang (2019) for the variation in visual imagery across frames, operationalized as a continuous metric of the change of pixels.

Step 6: Text mining for focal variables and project controls. We applied text mining and NLP to construct variables characterizing project outcomes and project controls from the 11,801 project web pages downloaded from Kickstarter. Prior crowdfunding studies have identified three key project outcomes: (1) total funding pledged (e.g., Burtch, Ghose, and Wattal 2016; Fan, Gao, and Steinhart 2020), (2) number of backers supporting the project (e.g., Fan, Gao, and Steinhart 2020; Younkin and Kuppuswamy 2018), and (3) project success (e.g., Fan, Gao, and Steinhart 2020; Li, Shi, and Wang 2019). We examined all three consequential dependent variables. From the project web pages, we extracted the first two measures and the project's funding goal to construct project success. A project is considered successful if the total amount of funding pledged meets or exceeds the project's stated funding goal (Mollick 2014).

In addition to the funding goal, we measured a set of project controls, as is typical in prior crowdfunding studies. To account for characteristics of the purchase options in a crowdfunding project, we constructed the menu length (the number of options) and the mean price of the options (Hu, Li, and Shi 2015). We included project duration (the length of time the project was active; Li, Shi, and Wang 2019). Finally, we measured the project's creator experience, which may impact the creator's ability to deliver an attractive product to consumers (Mukherjee, Chang, and Chattopadhyay 2019). The raw WebRobots data include information on all projects launched on Kickstarter since its inception, which we used to compute the number of prior projects by each creator. Table W3 in Web Appendix A lists all variable names, abbreviations (used in equations, described subsequently), and definitions.

Empirical Analyses and Results

We examined three consequential dependent variables that provide a comprehensive account of outcomes in online crowdfunding: (1) pledged funding, (2) number of backers, and (3) project success. We first present model-free evidence on the hypothesized effect of voice numerosity in Figure W1 (Web Appendix A) that depicts the means of these dependent variables across three levels of narrating voices: (1) one voice (45.38% of videos), (2) two voices (44.73% of videos), and (3) three or more voices (9.89% of videos). In line with our prediction, results show a monotonically increasing effect of the number of voices on all three dependent variables.

To formally test our predictions, our empirical specification relates each dependent variable to the number of narrator voices, the rate of delivery of the spoken message, and their interaction, which are the focal constructs based on our conceptual framework. We also included four sets of control variables (henceforth controls) in our empirical models: audial controls for differences in audial waveform, linguistic controls for differences in message tone, project controls for project-level differences (e.g., experience of the creator), and visual controls for visual elements.

Model specification and hypothesis testing. We estimate a Type I Tobit model for pledged amount (as the dependent variable is left-censored at 0):

[4] For example, project videos in our Kickstarter data set would generate more than 45 million images.

$$
\begin{aligned}
\text{pledged}_{p}^{*} = {} & \alpha_0 + \alpha_1 \times \text{num\_voices}_p + \alpha_2 \times \text{rate}_p + \alpha_3 \\
& \times \text{num\_voices}_p \times \text{rate}_p + \beta_1 \times \text{audial entropy}_p \\
& + \beta_2 \times \text{duration}_p + \beta_3 \times \text{energy}_p \\
& + \beta_4 \times \text{spectral\_centroid}_p + \beta_5 \times \text{spectral\_entropy}_p \\
& + \beta_6 \times \text{volume}_p + \beta_7 \times \text{zero}_p + \eta_1 \times \text{analytic}_p \\
& + \eta_2 \times \text{authentic}_p + \eta_3 \times \text{clout}_p + \eta_4 \times \text{tone}_p \\
& + \eta_5 \times \text{num\_words}_p + \delta_1 \times \text{creator\_experience}_p \\
& + \delta_2 \times \text{funding\_goal}_p + \delta_3 \times \text{menu\_length}_p \\
& + \delta_4 \times \text{price}_p + \delta_5 \times \text{proj\_duration}_p + \theta_1 \times \text{faces}_p \\
& + \theta_2 \times \text{scenes}_p + \theta_3 \times \text{visual\_variation}_p \\
& + \sum_{i=2}^{12} \gamma_{mi} \text{month}_{pi} + \sum_{j=2018}^{2019} \gamma_{yj} \text{year}_{pj} \\
& + \sum_{k=2}^{31} \gamma_{ck} \text{category}_{pk} + \varepsilon_p, \qquad (1)
\end{aligned}
$$

where pledged_p^* is observed if $\text{pledged}_p^* \geq 0$ and 0 otherwise; α_1 measures the impact of number of voices on the pledged amount in project p; controls for audial, linguistic, project, and visual elements relate to βs, ηs, δs, and θs, respectively; the fixed effects $\{\ \{\gamma_{mi}\}_{i=2}^{12}, \{\gamma_{yj}\}_{j=2018}^{2019}, \{\gamma_{ck}\}_{k=2}^{31}\}$ account for seasonal, annual, and category-specific differences across videos; and ε_p is the error term. We estimated analogous models to that of Equation 1 for other dependent measures: a Tobit model for number of backers (as it is left-censored at zero) and a Probit regression model for project success (as it is a binary outcome), where we observe success_p if $\text{pledged}_p \geq \text{goal}_p$.

Table 1 presents our results. Across all three dependent variables, we find that having more voices narrate the project message improves project outcomes (all α_1s > 0, all *p*s < .01). The effect is both statistically significant and economically important: having an additional narrator voice, ceteris paribus, is associated with $12,795 in additional funds raised, the support of 118 additional backers, and a 1.6% increase in the probability of project success.

Moreover, the effect is moderated by the rate at which the spoken content was delivered (all α_3s < 0, all *p*s < .05): having more voices narrate the project message at faster rates relates to *lowered* project outcomes, consistent with the interpretation that cognitive processing underlies the effect of voice numerosity. To further investigate the nature of this interaction, we examine the marginal effect of number of narrating voices on the pledged amount, number of backers, and project success. Figure W2 (in Web Appendix A) illustrates how the effect (as estimated using the models in Table 1) of having an additional voice is qualified by speech rate (i.e., message comprehension) in the video. Across all three project outcomes, results show that the benefit of an additional voice is higher for easier-to-comprehend videos (slower rates; e.g., one word said per second) than for more complex videos (faster rates; e.g., three words said per second).

Control variables. Results in Table 1 show that videos with higher volume raised more funding (β_6s > 0, *p*s < .001; see Oksenberg, Coleman, and Cannell 1986; Packwood 1974). Videos with a more dynamic audial track (β_1s > 0, *p*s < .01), more verbal information (η_5s > 0, *p*s < .05), and more visual information (θ_2s > 0, *p*s < .001) also improved project outcomes. Projects in which creators had more experience (δ_1s > 0, *p*s < .001), those with higher funding goals (δ_2s > 0, *p*s < .001), and those that offered more purchase options (δ_3s > 0, *p*s < .001) all were related to better project outcomes.

Sensitivity analyses. To test the stability of our findings, we conducted 18 sensitivity analyses across the dependent measures; in every case, our conclusions remain robust. The first set of analyses concerns funding goal for each dependent measure. We dropped the following sets of projects sequentially and reestimated the models: (1) projects with a fundraising goal of less than $1,000, (2) projects with a fundraising goal three (or more) standard deviations from the mean, and (3) projects with a fundraising goal greater than $1,000,000 (see Tables W4, W5, and W6 in Web Appendix A). Next, to test the stability of results with respect to our independent variable, number of narrating voices, we dropped projects with a number of voices three (or more) standard deviations from the mean and reestimated the models (see Table W7 in Web Appendix A).

To see if nature of the speech differs when there are more voices, we examined in each waveform with two or more voices (1) the proportion of content spoken by all nondominant voice(s) (M = 33%, median = 34%) and (2) the average number of exchanges[5] between different voices (M = 2.7, median = 2). We included these as speech controls and reestimated the models; results are shown in Table W8 in Web Appendix A. Finally, we used propensity score matching to create a matched sample with a "treatment" group (multiple voices) and a "control" group (a single voice) based on observed characteristics (Rosenbaum 2020) and reestimated the models. Table W9 in Web Appendix A describes the findings. Our conclusions remain robust in all 18 sensitivity analyses across all three dependent measures.

Validation study. We randomly selected 300 videos (4.6% of the sample) from our Kickstarter data for the validation analysis. A total of 905 U.S. participants on CloudResearch (43.3% women, 56.4% men, .3% prefer not to say; M_{age} = 38.8 years) coded the focal variable (number of voices) on this random subset of our data. Each participant was given two randomly selected videos (from this subset of 300) and was asked to code the number of narrator voices they heard in each product video.

Participants' coding exhibited high agreement on the number of narrator voices in the Kickstarter videos, with intraclass correlation coefficient of .91 (95% confidence interval [CI] = [.90, .93]; Shrout and Fleiss 1979). The human-

[5] For example, a waveform exhibiting the speech pattern "Voice A, Voice B, Voice A" would signify one exchange.

Table 1. Voice Numerosity and Speech Rate in Crowdfunding (Study 1).

	DV: Pledged Amount (in USD)	**DV: Number of Backers**	**DV: Project Success**
	(1)	**(2)**	**(3)**
(Intercept)	−1,426,847.00*	−12,566.02	−13.01
	(722,293.40)	(6,751.72)	(14.29)
Number of voices	12,794.71***	117.85***	.25***
	(3,708.29)	(34.88)	(.07)
Rate	4,381.49	58.98*	.13*
	(2,761.26)	(25.89)	(.05)
Number of voices × rate	−3,909.07*	−36.65*	−.09**
	(1,577.97)	(14.83)	(.03)
Audial controls			
Audial entropy	38,972.28***	394.61***	.46**
	(7,203.95)	(67.58)	(.14)
Duration	−40.06	−.36	−.001
	(26.45)	(.25)	(.001)
Energy	−7.70	−.08	−.0002
	(4.94)	(.05)	(.0001)
Spectral centroid	2,675,876.00	23,031.22	23.54
	(1,445,345.00)	(13,510.69)	(28.60)
Spectral entropy	11,352.80	94.56	.19
	(6,316.95)	(59.29)	(.12)
Volume	1,176.34***	11.58***	.03***
	(349.02)	(3.27)	(.01)
Zero crossings	3.05	.03	−.0000
	(5.54)	(.05)	(.0001)
Linguistic controls			
Analytical thinking	114.19***	.92**	.001
	(31.68)	(.30)	(.001)
Authenticity	37.50	.32	.0003
	(34.56)	(.32)	(.001)
Clout	139.90*	1.52**	−.001
	(57.93)	(.54)	(.001)
Emotional tone	−58.76*	−.87***	.001**
	(26.80)	(.25)	(.001)
Number of words	27.56**	.20*	.001*
	(10.65)	(.10)	(.0002)
Project controls			
Creator experience	3,590.89***	40.64***	.10***
	(206.64)	(1.93)	(.01)
Funding goal	.30***	.002***	—[a]
	(.02)	(.0001)	
Menu length	1,771.40***	16.03***	.04***
	(154.43)	(1.45)	(.003)
Price	−.26	−.10***	−.0002***
	(1.62)	(.02)	(.0000)
Project duration	−117.18	−1.83**	−.02***
	(67.10)	(.63)	(.001)
Visual controls			
Faces	−394.69	−2.22	−.02**
	(342.01)	(3.21)	(.01)
Number of scenes	724.48***	4.41***	.01***
	(43.70)	(.41)	(.001)
Visual variation	−8,440.36	145.66	.23
	(9,283.33)	(86.93)	(.18)

*$p < .05$.
**$p < .01$.
***$p < .001$ (all two-sided tests).
[a] The dependent variable (DV) in Model 3, project success, is constructed on the basis of the project's funding goal. Thus, funding goal is not included as an explanatory variable in this model.

coded measure and the machine-coded measure are similar; they differ by an average of .46 voices and a median of zero voices. The results validate the machine-coded measure of number of voices. Moreover, we report an ancillary analysis in Web Appendix B, which showed that the hypothesized effects in the main study (with the full sample) replicated across all three project outcomes in the validation data (the subset of 300).

Discussion

This study investigates the hypothesized voice numerosity effect in a real-world setting. Results show that on Kickstarter, having more voices narrate a product message is associated with improved project outcomes. The findings replicate across all three consequential dependent measures that are important in crowdfunding in the three largest supracategories on Kickstarter (spanning 31 categories, accounting for almost 70% of all funds raised). The measured effect size is managerially significant: for each additional voice in the project video, the average project saw (1) an increase of about $12,795 in pledged amount (a 39% increase), (2) 118 more customers backing the project (a 38% increase), and (3) a 1.6% greater probability that the project is successfully funded (a 6.5% increase). Finally, the effect is moderated by speech rate across project outcomes: having different voices present a message at faster rates relates to lowered project outcomes. To the extent that faster speech rate impedes cognitive processing (Goldinger, Pisoni, and Logan 1991; Moore, Hausknecht, and Thamodaran 1986; Smith and Shaffer 1995), the findings are consistent with our conceptualization that cognitive processing underlies the voice numerosity effect. These findings are replicated in the validation study with 300 randomly selected videos, using the human-coded measure of the number of voices.

Study 2: Voice Numerosity in Advertising

The primary purpose of Study 2 was to extend the findings of Study 1 in three important ways. First, we sought to generalize our findings in crowdfunding to another important real-world context for marketing practice: advertising. Second, whereas the crowdfunding projects in Study 1 focus on new product innovations, the ads in this study primarily focus on existing products. Third, this study examined the effect in relation to consumers' perceived efficacy of ads, complementing Study 1's examination of behavioral outcomes and building on the exploratory work of Hussain et al. (2017). They collected a data set of video ads to assess advertising efficacy using computer vision, with a focus on algorithmically analyzing the visual rhetoric of images in ads. We built on their data set to investigate voice numerosity.

Research Setting and Data

We obtained the video ad data set from Hussain et al. (2017), which includes the YouTube URLs and human-annotated characteristics of 2,449 ads in English[6] from YouTube and an internet service provider. Hussain et al. trained Amazon Mechanical Turk (MTurk) participants to code each ad along four dimensions: (1) a score of ad effectiveness from 1 to 5 (with 5 being "Most effective"), which serves as the main dependent measure in their study as well as ours; (2) measures of whether an ad is exciting and whether an ad is funny; (3) measures of sentiment (how the ad emotionally affects viewers); and (4) measures of the ad topic (e.g., restaurants).

Each video, on each dimension, was coded by about five human annotators. As the Hussain et al. (2017) data set does not include the ads directly, we utilized the ads' YouTube URLs (which are in the data set) to download the ads. Some URLs no longer worked,[7] leaving 1,610 ads for all subsequent analyses. We processed the ads in a similar manner as in Study 1 to measure the same sets of audial, linguistic, and visual controls. In Web Appendix C, Figures W3 and W4 depict the parsing procedure and variable construction; Table W10 lists the variable names, abbreviations in the equations, and definitions; Tables W11 and W12 list the sentiments and ad topics coded by MTurk participants; and Table W13 reports the summary statistics.

The analysis method and procedure closely followed that of Study 1, except in three aspects. First, in Study 1, we examined the hypothesized effect on relevant behavioral outcomes in the crowdfunding context, such as funding pledged to a project. In this study, we examined the effect on consumers' perceived efficacy of ads. Second, in the crowdfunding data set, all videos had the same resolution (full high definition), as required by the Kickstarter platform. This is not the case in the video ad data set, so we added a visual control—the video resolution—to account for image detail. Third, Hussain et al. (2017) collected, and included in their study, a comprehensive range of sentiments and topics across ads, which they used as controls. We replaced the crowdfunding project controls from our empirical specification in Study 1 (e.g., menu length) with these ad controls. These rich annotations help us pinpoint the voice numerosity effect. Figures W3 and W4 in Web Appendix C illustrate the parsing procedure for Study 2.

Empirical Analyses and Results

We specify an ordinal regression with a similar specification as in Study 1:

[6] As Hussain et al. (2017) focused on visuals of the ads, the raw data set with 3,477 video URLs included ads that were not in English. In the raw data set, among the 2,818 ads that were in English, 369 ads did not have human-annotated characteristics along the four dimensions.

[7] We wrote a computer script to download the video ads using the URLs provided, in January 2021. Subsequently, research assistants manually checked all URLs in Hussain et al.'s (2017) data set. Among the video ads that were no longer available, 68% indicated "Private video (not accessible)" (571); 31%, "Video unavailable" (261); and 1%, "This video has been removed" (7).

$$
\begin{aligned}
\text{effective}_a = {} & \alpha_0 + \alpha_1 \times \text{num\_voices}_a + \alpha_2 \times \text{rate}_a + \alpha_3 \\
& \times \text{num\_voices}_a \times \text{rate}_a + \delta_1 \times \text{exciting}_a + \delta_2 \\
& \times \text{funny}_a + \beta_1 \times \text{audial entropy}_a + \beta_2 \times \text{duration}_a \\
& + \beta_3 \times \text{energy}_a + \beta_4 \times \text{spectral\_centroid}_a + \beta_5 \\
& \times \text{spectral\_entropy}_a + \beta_6 \times \text{volume}_a + \beta_7 \times \text{zero}_a \\
& + \eta_1 \times \text{analytic}_a + \eta_2 \times \text{authentic}_a + \eta_3 \times \text{clout}_a \\
& + \eta_4 \times \text{tone}_a + \eta_5 \times \text{num\_words}_a + \theta_1 \times \text{faces}_a \\
& + \theta_2 \times \text{scenes}_a + \theta_3 \times \text{visual\_variation}_a \\
& + \theta_4 \times \text{resolution}_a + \sum_{i=2}^{18} \delta_{si} \, \text{sentiment}_a \\
& + \sum_{i=2}^{29} \delta_{tj} \, \text{topic}_{aj} + \varepsilon_a, \qquad (2)
\end{aligned}
$$

where α_1 measures the impact of number of narrator voices on scores of ad effectiveness in ad a; controls for advertising, audial, linguistic, and visual elements correspond to δs, βs, ηs, and θs, respectively; and ε_a is the error term. To ensure that our results are not sensitive to inclusion of the ad (sentiment and topic) controls (which have many levels) in the Hussain et al. (2017) data set, we estimate three models in which we sequentially introduced the sentiment and topic controls.

As shown in Table 2, we find that having more voices narrate an ad message significantly increases perceived ad efficacy (all α_1s > 0, all *p*s < .001), indicating a robust effect of voice numerosity across all three models. The effect is moderated by the rate at which the spoken ad message was delivered (all α_3s < 0, all *p*s < .01): having more voices narrate an ad message at faster rates relates to lower perceived ad efficacy. Further results of the marginal effect of narrating voice show that the benefit of an additional voice is greater for easier-to-comprehend ads (where ad messages are narrated at slower rates) than for more complex ads (ad messages narrated at faster rates; see Figure W5 in Web Appendix C).

Ad effectiveness increases for ads that are exciting (δ_1s > 0, *p*s < .001) and ads with a more dynamic audial track (β_1s > 0, *p*s < .05), longer duration (β_2s > 0, *p*s < .001), brighter sounds (β_4s > 0, *p*s < .001), and more visual variation (θ_3s > 0, *p*s < .001; see Table 2). We conducted robustness checks in which we reestimated the models after we (1) dropped ads with number of voices three (or more) standard deviations from the mean and (2) used propensity score matching to create a matched sample with a pseudo "treatment" group (multiple voices) and a "control" group (a single voice). Tables W14 and W15 in Web Appendix C show that our conclusions remain robust in these sensitivity analyses.

Table 2. Main Results in Ad Data Set (Study 2).

	DV: Ad Effectiveness		
	(1)	(2)	(3)
Number of voices	.248***	.303***	.289***
	(.039)	(.045)	(.051)
Rate	−.039	.003	.056
	(.051)	(.053)	(.056)
Number of voices × rate	−.066**	−.087***	−.083**
	(.023)	(.026)	(.028)
Ad controls			
Exciting	.377**	.533***	.640***
	(.120)	(.124)	(.126)
Funny	−.066	.103	.228*
	(.087)	(.114)	(.116)
Audial controls			
Audial entropy	.091*	.094*	.235***
	(.041)	(.043)	(.046)
Duration	4.670***	3.018***	3.985***
	(.001)	(.001)	(.002)
Energy	−.017	−.018*	−.025**
	(.009)	(.009)	(.009)
Spectral centroid	12.367***	11.362***	12.181***
	(.009)	(.009)	(.010)
Spectral entropy	−.006	−.006	.037
	(.025)	(.034)	(.038)
Volume	−4.452***	−2.208***	2.781***
	(.001)	(.001)	(.001)
Zero crossings	−.007	−.004	−.008
	(.005)	(.005)	(.005)
Linguistic controls			
Analytical thinking	−.001	−.001	−.0001
	(.001)	(.001)	(.001)
Authenticity	−.0002	.0001	−.0003
	(.001)	(.001)	(.001)
Clout	.001	.001	.0005
	(.001)	(.001)	(.001)
Emotional tone	.0005	.001	.001
	(.001)	(.001)	(.001)
Number of words	.013	.014	.010
	(.009)	(.009)	(.009)
Visual controls			
Faces	−.008	−.008	−.009
	(.010)	(.011)	(.011)
Number of scenes	−.001	−.001	−.0002
	(.002)	(.003)	(.003)
Visual variation	.956***	.969***	.812***
	(.003)	(.004)	(.008)
Resolution	−.00001	.00003	.00002
	(.00003)	(.00004)	(.00004)
Sentiments		Yes	Yes
Topics			Yes

**p* < .05.
***p* < .01.
****p* < .001 (all two-sided tests).

Discussion

The findings of this study generalize the hypothesized effect of voice numerosity to video advertising. The results are consistent with the interpretation that voice numerosity is facilitated by the heightened opportunity to process the spoken ad message, providing support for the role of cognitive processing through which the voice numerosity effect manifests on perceived ad efficacy. In the studies that follow, we test the hypothesized effect in controlled experiments.

Study 3: Voice Numerosity Under Varied Distraction

Our field studies established the voice numerosity effect in different real-world settings (crowdfunding and advertising) across product categories (e.g., games, automobiles, everyday products) and outcome measures (e.g., project success, perceived ad efficacy). In Study 3, we experimentally controlled for all elements of the stimulus and varied only voice numerosity, enabling us to rule out alternative accounts and to demonstrate the causal role of voice numerosity on consumer decisions. We also aimed to provide further evidence that the hypothesized effect is driven by enhanced attention and processing. When consumers have limited processing capacity, they may not be able to attend to and process a persuasive message, even if they detect the change in voice. As prior research shows, consumers suffer such processing deficits due to distraction (e.g., Nowlis and Shiv 2005) or time pressure (e.g., Siemer and Reisenzein 1998) when completing a focal task. Thus, we expected the voice numerosity effect to be moderated by processing ability.

Participants viewed a video about a new product and stated their maximum WTP for it. Half of the participants watched a video with one voice describing the product; the other half watched the same video with five different voices sequentially narrating. We varied participants' processing resources through a distraction task, directly manipulating their ability to pay attention to the focal task (see Spencer, Zanna, and Fong [2005] for a discussion of the "moderation-of-process" strategy to assess the underlying process). We predicted that participants' WTP for the product would be higher when the video is voiced by more narrators under high processing capacity (i.e., when participants are less distracted) than under low processing capacity (i.e., when participants are more distracted).

Method

Participants and design. A total of 382 U.S. participants (52% women, 48% men; $M_{age} = 34.4$ years) were recruited from the CloudResearch online panel for a small monetary compensation. They were randomly assigned to one of four conditions of a 2 (distraction: high vs. low) × 2 (number of voices: 1 vs. 5) between-subjects design.

Pretest. The purpose of the pretest was to test the effectiveness of the number-of-voices manipulation. We recruited 100 U.S. participants (44% women, 56% men; $M_{age} = 35.3$ years) from the CloudResearch online panel for a small monetary compensation. They were randomly assigned to one of two experimental conditions (number of voices: 1 vs. 5) in a between-subjects design. All were asked to watch a short video clip about a wireless charger and to count the number of voices heard.

To vary the number of voices, we created different versions of the video in which the content was identical across the versions, the only difference being the voice(s) narrating the spoken content. In the one-voice condition, the same voice conveys the entire product message in the voice-over. In the five-voice condition, five different voices convey the product message, each voice narrating a different portion in succession. We used five voice-synthesis models (which convert text to speech) to create the voices, counterbalanced using a Latin-square design to ensure comparability. We created modified videos by combining audio tracks from the voice-synthesis models with visual frames of the original video. We describe details of the video stimuli and results of a separate pretest on the perceived vocal qualities of the synthetic voices in Web Appendix D.

Results of the pretest showed that participants were fairly accurate in identifying the different number of spoken voices in the video. The average estimate was 1.63 in the one-voice condition and 4.58 in the five-voice condition ($F(1, 98) = 96.04$, $p < .0001$, $\eta^2 = .49$). Further comparison showed a significant difference in the median number of spoken voices across the experimental conditions, with a median of 1 in the one-voice condition and a median of 5 in the five-voice condition (Kruskal–Wallis test: $\chi^2(1) = 58.64$, $p < .0001$, $\varepsilon^2 = .59$).

Procedure and measures. The main experiment was administered as two "unrelated" studies. In the "first" study, under the pretense of a temporary memory test, participants were told that the study was interested in people's ability to remember unfamiliar numbers for a short period of time. We used a well-established operationalization to vary distraction (adapted from Nowlis and Shiv [2005], Experiment 1). In the high-distraction condition, participants were asked to memorize a ten-digit number; in the low-distraction condition, they were asked to memorize a two-digit number. They were reminded to hold the number in their memory and not write it down. Participants were informed that they would be asked to recall the number after a short delay and asked to complete another study in the meantime.

In the "second" study, participants were given a product-evaluation task in which they imagined that they were looking for a wireless charger for their cell phone and came across a new product in an online marketplace (Kickstarter). All watched a short video clip about the product. We varied the number of narrator voices describing the product while holding the visual and spoken content constant (same as the pretest video).

As the main dependent measure, participants indicated the maximum they were willing to pay for the product, reporting their WTP on a nine-point scale from $30 to $70 (presented in $5 increments). They then were asked to describe how they made their decision; this open-ended measure aimed to assess whether participants knowingly used the number of voices in their decision calculus. As a demand check, they were asked to guess the study's purpose.

To check for possible confounding effects of task involvement and mood, the former was assessed on three items (e.g., "I took the task of evaluating the product very seriously"; 1 = "Strongly disagree," and 7 = "Strongly agree"; $\alpha = .79$). Mood was measured on four seven-point items (e.g., "unpleasant"/"pleasant," "unhappy"/"happy"; $\alpha = .91$). As a manipulation check for number of voices in the video, participants selected a number from one to seven.

Participants then completed the second part of the temporary memory test. As manipulation checks for distraction, they were asked to indicate (1) the number they memorized and (2) how easy or difficult it was to remember (1 = "Very easy," and 7 = "Very difficult"). Finally, they reported background information such as gender, age, and general interest in product innovations on crowdfunding platforms (1 = "Not at all interested," and 7 = "Very interested").

Results

Preliminary analyses. No participant correctly guessed the purpose of the study, but 25 were removed for indicating they were generally "not at all interested" in product innovations on crowdfunding platforms (scoring 1 on a seven-point scale of reported general interest). Subsequent analyses were thus based on 357 observations. Participants noticed the different number of narrating voices. The average estimate was 1.20 in the one-voice condition and 2.50 in the five-voice condition ($F(1, 353) = 176.99$, $p < .0001$, $\eta^2 = .33$). Participants seemed aware of the spoken voices in the video even though they were not asked to pay attention to this detail. The number-of-voices manipulation was successful (see Web Appendix D for a summary of results on the effectiveness of the number-of-voices manipulation across studies).

To test the effectiveness of the distraction manipulation, an analysis of variance (ANOVA) of the reported ease or difficulty of remembering the number revealed only a main effect of distraction ($F(1, 353) = 795.99$, $p < .0001$, $\eta^2 = .69$), indicating greater difficulty remembering a ten-digit number ($M_{\text{high-distraction}} = 5.50$) than a two-digit number ($M_{\text{low-distraction}} = 1.48$). The accuracy of their recall reflected a similar pattern: Only 28% of participants in the high-distraction condition accurately recalled the number, compared with 98.35% in the low-distraction condition (Fisher's exact test, $p < .0001$). No differences in mood (ps > .25) or task involvement (ps > .22) were found across conditions.

Willingness to pay. An ANOVA of participants' WTP for the product yielded a marginally significant main effect of distraction ($F(1, 353) = 2.83$, $p = .093$, $\eta^2 = .008$). Participants were willing to pay slightly more in the low-distraction condition (M = \$45.93) than in the high-distraction condition (M = \$44.03). More importantly, participants' WTP exhibited a significant interaction of number of voices × distraction ($F(1, 353) = 4.90$, $p = .027$, $\eta^2 = .014$). As shown in Figure 2, with high distraction (less processing resources), participants' WTP was comparable whether the product was described by one voice (M = \$44.81) or five voices (M = \$43.40; F < 1). However, with low distraction (more processing resources), participants' WTP for the product was significantly higher with voice-over narration by more voices ($M_5 = \$47.98$ vs. $M_1 = \$44.18$; $F(1, 353) = 5.34$, $p = .021$, $\eta^2 = .015$), demonstrating the voice numerosity effect. Further, participants in the five-voice–low-distraction condition reported higher WTP (M = \$47.98) than participants in any of the other three conditions ($M_{\text{pooled}} = \$44.10$; $F(1, 353) = 7.78$, $p = .006$, $\eta^2 = .022$).

Additional analyses indicated that across the number-of-voices conditions, neither the voice itself ($F(4, 353) = 1.80$, $p = .13$) nor the ordering of the different voices ($F(4, 353) < 1$) influenced participants' WTP for the product. Ancillary analyses on participants' descriptions of their decision-making process are discussed in Web Appendix E.

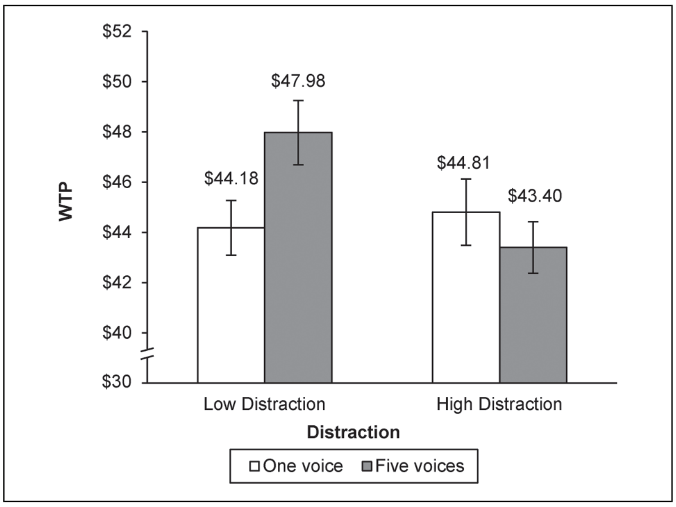

Figure 2. Voice Numerosity Moderated by Task Distraction (Study 3).
Notes: Error bars represent ±1 standard error.

Discussion

The results show that the voice numerosity effect can foster persuasion, depending on consumers' processing ability. With more processing resources (low distraction), participants' WTP for the target product was greater after watching a video wherein the voice-over had five different voices than after watching a video with a single voice. But with limited processing resources (high distraction), their WTP was comparable, irrespective of number of voices. These results were replicated in a lab experiment with 72 university students (see Web Appendix F). Due to the size of the participant pool that semester, we were unable to recruit more students and thus conducted the experiment using an online panel (reported as the main experiment).

Study 4: Voice Numerosity and Measured Cognitive Responses

Studies 1 to 3 documented our hypothesized effect of voice numerosity and showed that it was moderated by speech rate and distraction. In particular, the persuasive effect of multiple narrating voices is diminished when consumers' attention and processing is hindered (under faster speech rate and higher distraction). The purpose of Study 4 was to measure consumers' cognitive responses to directly test the process (Cacioppo and Petty 1981). If the effect is due to increased attention and processing of the product message when it is narrated by more voices, then the effect should be mediated by the favorability of participants' cognitive responses toward the product. The study also extended our investigation of the effect to consumers' purchase likelihood and to another product category.

Method

Participants and design. The study was conducted among U.S. participants from the Prolific panel who were prescreened on their general interest in product innovations on crowdfunding platforms (the same exclusion measure from Study 3). This helps ensure that the decision scenario would be relevant to participants. A total of 191 participants (58.1% women, 38.2% men, 3.1% nonbinary, .5% prefer not to say; M_{age} = 37.2 years) who qualified (scoring more than 1 on a seven-point scale of general interest) completed the study for a small monetary compensation. They were randomly assigned to one of the two experimental conditions (number of voices: one vs. five; https://aspredicted.org/zj87u.pdf).

Procedure and measures. Participants completed a product-evaluation task that was identical to Study 3 except in three aspects. First, the target product was a smart mug that keeps hot beverages at a certain temperature. Second, as the main dependent measure, participants indicated their purchase likelihood on a single item (1 = "Definitely will not buy it," and 7 = "Definitely will buy it"). Third, participants were asked to write the thoughts they had as they watched the video (adapted from Cacioppo and Petty 1981). On the ensuing screen, we displayed the thoughts that participants had written and asked them to code each thought as negative, neutral, positive, or irrelevant to the product and/or the video clip. Because we theorized that the effect is due to increased attention and processing of the product message when it is narrated by more voices, we also asked participants to code each thought by its type, that is, whether the thought was about the product, the video, both, or neither (Chattopadhyay and Basu 1990). This allowed us to assess the effect of voice numerosity on product-related responses, compared with prior speculation that voice drives persuasion through greater focus on the narrator but less focus on the product (Chaiken and Eagly 1983; Grewal, Gupta, and Hamilton 2021). The same manipulation check (of number of voices), demand check, and confounding checks of involvement ($\alpha = .72$) and mood ($\alpha = .93$) were collected. Finally, participants reported basic background information (gender, age) and whether the video and audio track loaded properly for the product-evaluation task, as in Study 3. We describe the study in detail in Web Appendix G.

Results and Discussion

Preliminary analyses. Four participants closely guessed the hypothesis (e.g., "how effectively one can be persuaded to purchase a product from a Kickstarter promotional video which uses AI voice"); their data were removed. All analyses were based on 187 observations. The number-of-voices manipulation was successful ($M_1 = 1.05$ vs. $M_5 = 2.32$; $t(185) = 11.04$, $p < .0001$, $\eta^2 = .40$). Participants were comparable in task involvement ($F < 1$) and mood ($F < 1$).

Purchase likelihood. An ANOVA of participants' purchase likelihood for the product yielded a significant effect of number of voices ($t(185) = 2.17$, $p = .031$, $\eta^2 = .025$) such that narration by more voices increased participants' purchase likelihood ($M_5 = 4.17$ vs. $M_1 = 3.57$). The counterbalancing was effective: neither the voice (one-voice conditions; $F < 1$) nor the order of voices (five-voice conditions; $t(185) = 1.40$, $p > .16$) affected participants' purchase likelihood.

Cognitive responses. Prior research on the elaboration likelihood model has indicated that greater message processing may affect the quality (valence) but not the quantity (number) of thoughts listed (Petty and Cacioppo 1986). Thus, following prior research (Cacioppo and Petty 1981; Chattopadhyay and Basu 1990), we used participants' coding to compute the proportion of favorable product-related cognitive responses (the extent of positive responses, divided by the sum of positive and negative responses). Higher values reflected a greater proportion of favorable cognitive responses. Analysis of participants' cognitive responses showed a significant effect of number of voices ($t(185) = 2.09$, $p = .038$, $\eta^2 = .02$). Participants' thoughts were more favorable toward the product when the product message was narrated by more voices ($M_5 = .60$ vs. $M_1 = .47$).

Mediation. We conducted a bias-corrected mediation model (PROCESS version 4.0, Model 4), with 10,000 bootstrap samples as recommended by Hayes (2022). As shown in Figure 3 and consistent with our theorizing, cognitive responses significantly mediated the effect of number of narrating voices (0 = one voice, 1 = five voices) on purchase likelihood, as the 95% CI for the indirect effect excluded 0 (95% CI = [.0051, .6162]). The effect of number of voices became nonsignificant when the mediator was added ($\beta = .30$, SE = .24, $p = .21$).

In summary, the findings show that the effect of number of voices on purchase likelihood is mediated by favorability of cognitive responses toward the product. The study extends the findings in Studies 1 to 3 by providing direct evidence for the role of product-related thoughts in driving the effect of voice numerosity on purchase likelihood.

General Discussion

Marketers seek to maximize consumers' attention to and processing of messages in product videos and broadcast ads to increase effectiveness of their communication. The present research shows that persuasive power of a video can be enhanced when its spoken narration employs more voices. Specifically, results from four studies (plus validation and replication studies reported in the Web Appendices)—including real-world data sets on crowdfunding and video advertising—provide consistent evidence that voice numerosity can improve consequential crowdfunding project outcomes (pledged amount, number of backers, and project success; Study 1), perceived advertising efficacy (Study 2), consumers' WTP for the target product (Study 3), and consumers' purchase likelihood (Study 4). The voice numerosity effect is moderated by the rate of narration of the message (Studies 1 and 2) and consumers' processing resources (Study 3); these moderators relate to the opportunity and ability to process a persuasive

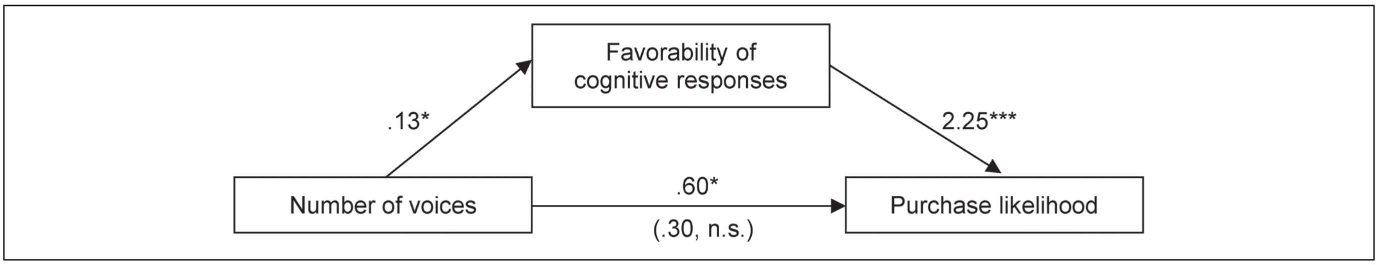

Figure 3. Voice Numerosity Mediated by Cognitive Responses (Study 4).
$*p < .05$.
$**p < .01$.
$***p < .001$.
Notes: The path coefficients are unstandardized betas. Value in parentheses is the effect of the independent variable on the DV after controlling for the mediator.

message (see MacInnis and Jaworski 1989). Moreover, the effect is mediated by the favorability of cognitive responses toward the product (Study 4). Testing two theoretically derived boundaries of voice numerosity and measuring people's cognitive responses provided evidence for our conceptualization that consumers' attention and processing underlie the voice numerosity effect.

The effect emerged in a wide array of categories, including 31 product categories on Kickstarter (under the Design, Games, and Technology supracategories; Study 1) and 29 topics in online ads (e.g., cars, clothing, environment; Study 2). The effect also generalized across human voices (Studies 1 and 2) and synthetic voices (Studies 3 and 4) conveying the message. We observed the voice numerosity effect across diverse samples (consumers, MTurk participants, and students) and settings (real-world marketplaces and controlled experiments). The broad range of operationalizations of voice stimuli, product categories, types of video marketing, consumer samples, and consumption behavior shows the potential generalizability of the voice numerosity effect.

These results contribute to the consumer behavior literature in multiple ways. First, we add to the literature on the effect of voice on consumer behavior (Dahl 2010), which has received little research attention to date (Krishna and Schwarz 2014). We concur with prior research (e.g., Chattopadhyay et al. 2003; Wang et al. 2021) that narrator voices can exert powerful effects on consumer behavior. We add to the previous findings by showing that the number of voices in video narration can affect consumer behavior—despite myriad sensory signals—through the central route of persuasion.

Second, we add to the literature on persuasive marketing communications. Prior research has uncovered various elements that aid the design of communications through message content or communicator characteristics (see Petty and Cacioppo 1986), placing more emphasis on the effect of visual features on consumer processing. Less research has examined the manner in which the message is delivered, including "the fascinating question of what roles audition may play in marketing and the way that consumers process information" (Meyers-Levy, Bublitz, and Peracchio 2010, p. 138). Our research addresses this question and responds to recent calls to study the effect of the evolving information environment on consumer behavior (Simonson 2015), as well as to employ advances in research tool kits to answer consumer-relevant questions (Inman et al. 2018). Voice and multimedia are ubiquitous in marketing, but research has been limited by methodological challenges in designing these stimuli for experimental research (see Krishna and Schwarz 2014) and analyzing unstructured multimedia data for empirical modeling (see Grewal 2018). We demonstrate the use of machine learning and NLP to overcome these challenges. These tools allow us to (1) examine consequential dependent variables at scale, leveraging new secondary data sources such as crowdfunding data (Study 1) and online video advertising data (Study 2), and (2) create multimedia materials for experiments, such as synthetic voices in Studies 3 and 4.

Theoretical Elaborations and Directions for Future Research

Intranarrator voice numerosity. As we aimed to provide an initial test for the persuasive effect of voice numerosity, we focused on voice variation among different narrators, in line with most prior studies using spoken voice stimuli (e.g., Goldinger, Pisoni, and Logan 1991). All else equal, our natural voices are likely to vary more between narrators than within a narrator, due to (1) the physical anatomy of narrators' vocal tracts producing sounds (Fant 1960), (2) their natural idiosyncrasies in how they speak (Newman, Clouse, and Burnham 2001), and (3) indexical information on their identity (e.g., age) and emotional state (Belin, Fecteau, and Bédard 2004). We believe what matters is message recipients' subjective experience upon hearing the voices conveying the message. It is possible for listeners to perceive the same narrator's voice to be different (such as when a voice-over artist intentionally modulates their voice); future research can assess whether (and to what extent) voice changes within the same narrator would produce the effect.

Numerosity in other auditory and visual dimensions. One interesting question is whether numerosity (or variation) in other auditory or visual dimensions, among myriad sensory signals, can generate a similar effect as voice numerosity. We suspect that numerosity (or stimulus change) in other dimensions might *not* necessarily lead to message recipients' increased attention and processing. This speculation is consistent with the literature on change blindness with visual stimuli (e.g., Rensink 2002; Simons and Levin 1997), wherein studies document that people often fail to notice even relatively large changes in their visual scenes. In contrast, a change in voice can involuntarily capture attention (Cherry 1953), perhaps due to the significance of vocal stimulus and voice-preferential processing (Charest et al. 2009). This speculation is consistent with our empirical results using two real-world data sets. Among the visual and audial variables, the findings show that only the number of voices exerts consistent and significant effects on all dependent variables (DVs) in all analyses across both data sets. In contrast, numerosity in visual variables (e.g., number of scenes, visual variation, presence of faces) did not always replicate across analyses, nor did numerosity in other audial variables (e.g., zero crossings, spectral entropy). Thus, while we suspect that the effect might not necessarily generalize to other forms of numerosity, we believe the findings present fruitful directions for future research to circumscribe the conditions in which numerosity in other audial and visual dimensions influences consumer behavior.

Voice numerosity and other audial and visual dimensions. Another interesting direction is whether voice numerosity might interact with other dimensions such as image and/or text, akin to Li and Xie's (2020) study of fit effects between image and text in social media posts. Although they did not find image–text fit effects consistently across their Twitter and Instagram data sets, future research can develop new theoretical frameworks to explore potential fit effects between voice and other audial/visual dimensions in today's media-rich environment. Would these image features interact with vocal features? Research is needed to explore these interesting questions.

Auditory attention. One limitation of the present research is that, although we theorize increased attention from changes in voices narrating a product message, we did not directly measure auditory attention. Unlike visual attention, auditory attention is mostly independent of the position of the head and ears (Scharf 1998). Thus, auditory attention is often identified by measuring electrophysiological data such as electroencephalograms (Alickovic et al. 2019). We believe more systematic research is warranted to obtain further evidence on the role of attention in voice numerosity. For example, future work can use neuroscience techniques to study how hearing different voices shapes consumer attention.

Relation to encoding variability. It is interesting to relate our findings to research on sensory stimuli and cognition (i.e., how recipients process stimuli). One such area is the broad memory literature on encoding variability (Rose 1980; Unnava and Burnkrant 1991) and its related hypotheses, including levels of processing (Craik and Lockhart 1972) and distinctiveness (Gallo et al. 2008; Hunt 2006). These hypotheses center on encoding and retrieval factors that affect memory performance. In particular, encoding variability holds that a stimulus (such as a word or an ad) that is variably encoded in repeated presentations can aid recall (Rose 1980; Unnava and Burnkrant 1991). The levels-of-processing framework adds that improved memory performance results from semantic encoding ("deeper" processing) of the stimulus, which strengthens the stimulus's memory trace (Craik and Lockhart 1972). The distinctiveness explanation posits that it does so by leading to more distinctive memory traces (Hunt 2006).

There are notable conceptual differences between these memory frameworks and our work. First, prior studies on encoding variability focused on the effect of *repeated* presentations (i.e., variation in encoding conditions) of a stimulus on *memory* (Rose 1980). In contrast, we focus on the *initial* presentation of the target stimulus on *persuasion.* Second, prior memory studies on levels of processing and distinctiveness centered on the stimuli's semantic content (e.g., a word; Gallo et al. 2008), that is, the semantic characteristics (e.g., "deeper" processing of the meaning of a word; Craik and Lockhart 1972) or semantic elaboration (e.g., number of semantic features of a word; Hargreaves et al. 2012). In contrast, our research focuses on the effect of a nonsemantic element (voices) of the stimulus. In our experiments, the provided stimulus is identical in all semantic aspects across conditions.

Relation to prior work on encoding of perceptual and semantic information via audio. Prior memory studies have employed spoken-voice stimuli to study the encoding of perceptual and semantic information delivered by audio. Perhaps because these studies aimed to uncover the basic memory process, much shorter tokens were used as target stimuli (e.g., vowels, phonemes, words; Goldinger, Pisoni, and Logan 1991; Martin et al. 1989; Morton, Crowder, and Prussin 1971). Notably, this literature has shown that findings from studies using different forms of short tokens do not generalize when other forms of tokens are used. For example, results from word-list studies frequently do not hold for connected words or word pairs (Hunt and Einstein 1981; see Unnava and Burnkrant 1991). In contrast, persuasive marketing communications—the focus of our research—are typically much longer, in the form of meaningfully connected sentences (e.g., in our Studies 3 and 4, the spoken message is over 200 words). We thus add to prior work by showing that spoken narration, as a longer form of spoken-voice stimuli, can exert a persuasive influence on message recipients even in the presence of other visual and audial signals.

Relation to the multiple-source effect. It is interesting to consider how our research might relate to the effect of multiple social sources in persuasion. The multiple-source effect refers to the persuasive influence of having more social sources advocate a counterattitudinal position (Harkins and Petty 1981, 1987). In a typical study, student participants exhibited greater attitude change in favor of a senior comprehensive exam when they were explicitly informed that three students (vs. one student) were promoting the exam. The effect is due to greater perceived information utility with more advocates (i.e., social sources); it disappeared when participants were informed that the multiple advocates belonged to the same source (e.g., same academic committee; Harkins and Petty 1987). The multiple-source effect is thus unlikely to account for our results. First, in our studies, a persuasive message comes from only one source—the brand or firm behind the product in the video—as is typical in persuasive marketing communications. Second, in our real-world data sets, the number of faces that appeared in the videos (a visual cue for social sources) did not affect the outcome measures. Third, in our controlled experiments, the voice(s) conveying the message are machine-generated and not from humans (i.e., social others). Future research can vary social sources and voices orthogonally to assess whether they would have an additive or multiplicative effect on persuasion. Relatedly, does the video source moderate the voice numerosity effect on consumers' responsiveness to the product? We believe these are fruitful avenues, and we encourage future research to explore these questions.

Generalizability across various forms of communications. Although our research focuses on two prevalent forms of video marketing—product videos and video advertisements—we believe that the phenomenon would be observed in other forms of asynchronous communication. For example, a widespread (albeit different) form of asynchronous communication for companies is an earnings conference call, in which companies convey their financial results to interested parties such as investors, analysts, and the public, typically via webcast. In such a setting, would the voice numerosity effect manifest in earnings conference calls and investor behavior? Future research can explore whether the nature of communications matters for voice numerosity.

Managerial Implications

The Marketing Science Institute (2020) identified in its top research priorities the need for approaches "to capture and analyze non-structured data such as video, voice, and text in order to improve firm communications and customer experience" (p. 9). Our research speaks to this issue. As video marketing continues to affect consumers' purchase journeys, our research offers recommendations on voice-over narration for practitioners and architects of the consumer information environment (e.g., user experience designers) to consider in designing video communications. Current industry practice focuses on the need for "a clear speaking voice" (YouTube Advertising 2019) that signals "authority" and "relatability" (Voices 2018) but has yet to consider using number of voices as a strategic design element (which was also revealed in our interviews with senior executives). Our findings suggest that for more difficult-to-comprehend product messages (e.g., said at three words per second), it might be more effective to have just one narrator. In contrast, for messages that are simple to comprehend (e.g., said at one word per second), it may be worthwhile to have multiple narrators, to leverage the voice numerosity effect.

Author Note

The first two authors contributed equally.

Acknowledgments

The authors gratefully acknowledge the constructive input of the *JMR* review team, and thank Michelle P. Lee and seminar participants at Johns Hopkins University, the Behavioral Sciences Institute, Lingnan University, Northwestern University, and Peking University for helpful feedback.

Associate Editor

James R. Bettman

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The research was supported by the Singapore Ministry of Education (MOE) Academic Research Fund (AcRF) Tier 2 Grant, MOE2018-T2-1-181, to the first author.

ORCID iDs

Hannah H. Chang https://orcid.org/0000-0001-8653-0990
Anirban Mukherjee https://orcid.org/0000-0001-6381-814X

References

Aeschlimann, Mélanie, Jean-François Knebel, Micah M. Murray, and Stephanie Clarke (2008), "Emotional Pre-Eminence of Human Vocalizations," *Brain Topography*, 20 (4), 239–48.

Alickovic, Emina, Thomas Lunner, Fredrik Gustafsson, and Lennart Ljung (2019), "A Tutorial on Auditory Attention Identification Methods," *Frontiers in Neuroscience*, 13, 153.

Alpert, Judy I. and Mark I. Alpert (1990), "Music Influences on Mood and Purchase Intentions," *Psychology and Marketing*, 7 (June), 109–33.

Anand, Punam and Brian Sternthal (1990), "Ease of Message Processing as a Moderator of Repetition Effects in Advertising," *Journal of Marketing Research*, 27 (3), 345–53.

Apple, William, Lynn Streeter, and Robert Krauss (1979), "Effects of Pitch and Speech Rate on Personal Attributions," *Journal of Personality and Social Psychology*, 37 (5), 715–27.

Belin, Pascal, Shirley Fecteau, and Catherine Bédard (2004), "Thinking the Voice: Neural Correlates of Voice Perception," *Trends in Cognitive Sciences*, 8 (3), 129–35.

Belin, Pascal and Robert J. Zatorre (2003), "Adaptation to Speaker's Voice in Right Anterior Temporal Lobe," *NeuroReport*, 14 (16), 2105–09.

Brown, Bruce, William Strong, and Alvin Rencher (1974), "Perceptions of Personality from Speech: Effects of Manipulations of Acoustical Parameters," *Journal of the Acoustical Society of America*, 54 (1), 29–35.

Burtch, Gordon, Anindya Ghose, and Sunil Wattal (2016), "Secret Admirers: An Empirical Examination of Information Hiding and

Contribution Dynamics in Online Crowdfunding," *Information Systems Research*, 27 (3), 478–96.

Cacioppo, John T. and Richard E. Petty (1981), "Social Psychological Procedures for Cognitive Response Assessment: The Thought Listing Technique," in *Cognitive Assessment*, Thomas Merluzzi, Carol Glass, and Myles Genest, eds. Guilford Press, 309–42.

Calder, Bobby J., Chester A. Insko, and Ben Yandell (1974), "The Relation of Cognitive and Memorial Processes to Persuasion in a Simulated Jury Trial," *Journal of Applied Social Psychology*, 4 (1), 62–93.

Cavanaugh, Lisa A., James R. Bettman, and Mary Frances Luce (2015), "Feeling Love and Doing More for Distant Others: Specific Positive Emotions Differentially Affect Prosocial Consumption," *Journal of Marketing Research*, 52 (5), 657–73.

Chaiken, Shelly and Alice Eagly (1983), "Communication Modality as a Determinant of Persuasion: The Role of Communicator Salience," *Journal of Personality and Social Psychology*, 45 (2), 241–56.

Charest, Ian, Cyril R. Pernet, Guillaume A. Rousselet, Ileana Quiñones, Marianne Latinus, Sarah Fillion-Bilodeau, Jean-Pierre Chartrand, and Pascal Belin (2009), "Electrophysiological Evidence for an Early Processing of Human Voices," *BMC Neuroscience*, 10 (1), 127.

Chattopadhyay, Amitava and Kunal Basu (1990), "Humor in Advertising: The Moderating Role of Prior Brand Evaluation," *Journal of Marketing Research*, 27 (4), 466–76.

Chattopadhyay, Amitava, Darren W. Dahl, Robin J.B. Ritchie, and Kimary N. Shahin (2003), "Hearing Voices: The Impact of Announcer Speech Characteristics on Consumer Response to Broadcast Advertising," *Journal of Consumer Psychology*, 13 (3), 198–204.

Chelba, Ciprian, Tomas Mikolov, Mike Schuster, Qi Ge, and Thorsten Brants (2013), "One Billion Word Benchmark for Measuring Progress in Statistical Language Modeling," arXiv, https://doi.org/10.48550/arXiv.1312.3005.

Cherry, Colin E. (1953), "Some Experiments on the Recognition of Speech with One and with Two Ears," *Journal of the Acoustical Society of America*, 25 (5), 975–79.

Craik, Fergus and Kim Kirsner (1974), "The Effect of Speaker's Voice on Word Recognition," *Quarterly Journal of Experimental Psychology*, 26 (2), 274–84.

Craik, Fergus and Robert S. Lockhart (1972), "Levels of Processing: A Framework for Memory Research," *Journal of Verbal Learning and Verbal Behavior*, 11 (6), 671–84.

Cramer-Flood, Ethan (2021), "US Time Spent with Media 2021 Update," eMarketer (February 4), https://content-na1.emarketer.com/us-time-spent-with-media-2021-update.

Dahl, Darren W. (2010), "Understanding the Role of Spokesperson Voice in Broadcast Advertising," in *Sensory Marketing: Research on the Sensuality of Products*, Aradhna Krishna, ed. Routledge, 169–82.

Dhanani, Qahir and Anirban Mukherjee (2017), "Is Crowdfunding the Silver Bullet to Expanding Innovation in the Developing World?" World Bank Private Sector Development Blog (November 13), https://blogs.worldbank.org/psd/crowdfunding-silver-bullet-expanding-innovation-developing-world.

Edell, Julie A. and Marian Chapman Burke (1987), "The Power of Feelings in Understanding Advertising Effects," *Journal of Consumer Research*, 14 (3), 421–33.

Escera, Carles, Kimmo Alho, Istvan Winkler, and Risto Näätänen (1998), "Neural Mechanisms of Involuntary Attention to Acoustic Novelty and Change," *Journal of Cognitive Neuroscience*, 10 (5), 590–604.

Facebook IQ (2019), "How to Unlock Your Creative Potential on Stories," (March 19), https://www.facebook.com/business/news/insights/how-to-unlock-your-creative-potential-on-stories.

Fan, Tingting, Leilei Gao, and Yael Steinhart (2020), "The Small Predicts Large Effect in Crowdfunding," *Journal of Consumer Research*, 47 (4), 544–65.

Fant, Gunnar (1960), *Acoustic Theory of Speech Production*. Mouton.

Flaks, Jason, Shane Walker, Iroro Orife, and Morten Pedersen (2018), "Automatic Speech Recognition (ASR) Model Training," U.S. Patent 20180315417 A1.

Forehand, Mark and Andrew Perkins (2005), "Implicit Assimilation and Explicit Contrast: A Set/Reset Model of Response to Celebrity Voice-Overs," *Journal of Consumer Research*, 32 (3), 435–41.

Furedy, John J. and John Scull (1971), "Orienting-Reaction Theory and an Increase in the Human GSR Following Stimulus Change Which Is Unpredictable but Not Contrary to Prediction," *Journal of Experimental Psychology: General*, 88 (2), 292–94.

Gafni, Hadar, Dan Marom, and Orly Sade (2019), "Are the Life and Death of an Early-Stage Venture Indeed in the Power of the Tongue? Lessons from Online Crowdfunding Pitches," *Strategic Entrepreneurship Journal*, 13 (1), 3–23.

Gallo, David A., Nathaniel G. Meadow, Elizabeth L. Johnson, and Katherine T. Foster (2008), "Deep Levels of Processing Elicit a Distinctiveness Heuristic: Evidence from the Criterial Recollection Task," *Journal of Memory and Language*, 58 (4), 1095–11.

Gati, Itamar and Gershon Ben-Shakhar (1990), "Novelty and Significance in Orientation and Habituation: A Feature Matching Approach," *Journal of Experimental Psychology: General*, 119 (3), 251–63.

Goldinger, Stephen, David Pisoni, and John Logan (1991), "On the Nature of Talker Variability Effects on Recall of Spoken Word Lists," *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 17 (1), 152–62.

Graves, Alex, Abdel-rahman Mohamed, and Geoffrey Hinton (2013), "Speech Recognition with Deep Recurrent Neural Networks," 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, 6645–49, https://doi.org/10.1109/ICASSP.2013.6638947.

Grewal, Rajdeep (2018), "Marketing Insights from Multimedia Data: Text, Image, Audio, and Video," American Marketing Association (June 21), https://www.ama.org/2018/06/21/marketing-insights-from-multimedia-data-text-image-audio-and-video/.

Grewal, Rajdeep, Sachin Gupta, and Rebecca Hamilton (2021), "Marketing Insights from Multimedia Data: Text, Image, Audio, and Video," *Journal of Marketing Research*, 58 (6), 1025–33.

Grossmann, Tobias, Regine Oberecker, Stefan Paul Koch, and Angela D. Friederici (2010), "The Developmental Origins of Voice Processing in the Human Brain," *Neuron*, 65 (6), 852–58.

Hargreaves, Ian, Penny Pexman, Jeremy Johnson, and Lenka Zdrazilova (2012), "Richer Concepts Are Better Remembered: Number of Features Effects in Free Recall," *Frontiers in Human Neuroscience*, 6, 73.

Harkins, Stephen G. and Richard E. Petty (1981), "The Multiple Source Effect in Persuasion: The Effects of Distraction," *Personality and Social Psychology Bulletin*, 7 (4), 627–35.

Harkins, Stephen G. and Richard E. Petty (1987), "Information Utility and the Multiple Source Effect," *Journal of Personality and Social Psychology*, 52 (2), 260–68.

Hayes, Andrew F. (2022), *Introduction to Mediation, Moderation, and Conditional Process Analysis: A Regression-Based Approach,* 3rd ed. Guilford Press.

Hoffman, Donald D. and Manish Singh (1997), "Salience of Visual Parts," *Cognition*, 63 (1), 29–78.

Horowitz, Seth S. (2012), *The Universal Sense: How Hearing Shapes the Mind*. Bloomsbury.

Hu, Ming, Xi Li, and Mengze Shi (2015), "Product and Pricing Decisions in Crowdfunding," *Marketing Science*, 34 (3), 331–45.

Hunt, R. Reed (2006), "The Concept of Distinctiveness in Memory Research," in *Distinctiveness and Memory*, R. Reed Hunt and James Worthen, ed. Oxford University Press, 3–25.

Hunt, R. Reed and Gilles O. Einstein (1981), "Relational and Item-Specific Information in Memory," *Journal of Verbal Learning and Verbal Behavior*, 20 (5), 497–514.

Hussain, Zaeem, Mingda Zhang, Xiaozhong Zhang, Keren Ye, Christopher Thomas, Zuha Agha, Nathan Ong, and Adriana Kovashka (2017), "Automatic Understanding of Image and Video Advertisements," 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 1100–1110, https://doi.org/10.1109/CVPR.2017.123.

Inman, J. Jeffrey, Margaret C. Campbell, Amna Kirmani, and Linda L. Price (2018), "Our Vision for the *Journal of Consumer Research*: It's All About the Consumer," *Journal of Consumer Research*, 44 (5), 955–59.

Janiszewski, Chris, Andrew Kuo, and Nader T. Tavassoli (2013), "The Influence of Selective Attention and Inattention to Products on Subsequent Choice," *Journal of Consumer Research*, 39 (6), 1258–74.

Kahneman, Daniel (1973), *Attention and Effort*. Prentice-Hall.

Krishna, Aradhna and Norbert Schwarz (2014), "Sensory Marketing, Embodiment, and Grounded Cognition: A Review and Introduction," *Journal of Consumer Psychology*, 24 (2), 159–68.

Lehman, David, Kieran O'Connor, and Glenn R. Carroll (2019), "Acting on Authenticity: Individual Interpretations and Behavioral Responses," *Review of General Psychology*, 23 (1), 19–31.

Li, Xi, Mengze Shi, and Xin Wang (2019), "Video Mining: Measuring Visual Information Using Automatic Methods," *International Journal of Research in Marketing*, 36 (2), 216–31.

Li, Yiyi and Ying Xie (2020), "Is a Picture Worth a Thousand Words? An Empirical Study of Image Content and Social Media Engagement," *Journal of Marketing Research*, 57 (1), 1–19.

Liu, Xuan, Savannah Wei Shi, Thales Teixeira, and Michel Wedel (2018), "Video Content Marketing: The Making of Clips," *Journal of Marketing*, 82 (4), 86–101.

MacInnis, Deborah J. and Bernard J. Jaworski (1989), "Information Processing from Advertisements: Toward an Integrative Framework," *Journal of Marketing*, 53 (4), 1–23.

Mairesse, François, Marilyn A. Walker, Matthias R. Mehl, and Roger K. Moore (2007), "Using Linguistic Cues for the Automatic Recognition of Personality in Conversation and Text," *Journal of Artificial Intelligence Research*, 30, 457–500.

Makino, Takaki, Hank Liao, Yannis Assael, Brendan Shillingford, Basilio Garcia, Otavio Braga, and Olivier Siohan (2019), "Recurrent Neural Network Transducer for Audio-Visual Speech Recognition," arXiv, https://doi.org/10.48550/arXiv.1911.04890.

Marketing Science Institute (2020), "2020–2022 Research Priorities," (accessed May 13, 2021), https://www.msi.org/wp-content/uploads/2021/07/MSI-2020-22-Research-Priorities-final.pdf-WORD.pdf.

Martin, Christopher, John Mullennix, David Pisoni, and W. Van Summers (1989), "Effects of Talker Variability on Recall of Spoken Word Lists," *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 15 (4), 676–84.

Melumad, Shiri, J. Jeffrey Inman, and Michel Tuan Pham (2019), "Selectively Emotional: How Smartphone Use Changes User-Generated Content," *Journal of Marketing Research*, 56 (2), 259–75.

Meyers-Levy, Joan, Melissa G. Bublitz, and Laura A. Peracchio (2010), "The Sounds of the Marketplace: The Role of Audition in Marketing," in *Sensory Marketing: Research on the Sensuality of Products*, Aradhna Krishna, ed. Routledge, 137–56.

Millward Brown (2012), "How Should Voiceovers Be Used in Ads?" (March), http://www.brandreportblog.com/wp-content/uploads/MillwardBrown_KnowledgePoint_Voiceover.pdf.

Mollick, Ethan (2014), "The Dynamics of Crowdfunding: An Exploratory Study," *Journal of Business Venturing*, 29 (1), 1–16.

Moore, Danny, Douglas Hausknecht, and Kanchana Thamodaran (1986), "Time Compression, Response Opportunity, and Persuasion," *Journal of Consumer Research*, 13 (1), 85–99.

Morton, John, Robert G. Crowder, and Harvey A. Prussin (1971), "Experiments with the Stimulus Suffix Effect," *Journal of Experimental Psychology*, 91 (1), 169–90.

Mukherjee, Anirban, Hannah H. Chang, and Amitava Chattopadhyay (2019), "Crowdfunding: Sharing the Entrepreneurial Journey," in *Handbook of the Sharing Economy*, Russell W. Belk, Giana Eckhardt, and Fleura Bardhi, eds. Edward Elgar Publishing, 152–62.

Newman, Rochelle S., Sheryl A. Clouse, and Jessica L. Burnham (2001), "The Perceptual Consequences of Within-Talker Variability in Fricative Production," *Journal of the Acoustical Society of America*, 109 (3), 1181–96.

Nosofsky, Robert (1984), "Choice, Similarity, and Context Theory of Classification," *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 10 (1), 104–14.

Nowlis, Stephen and Baba Shiv (2005), "The Influence of Consumer Distractions on the Effectiveness of Food-Sampling Programs," *Journal of Marketing Research*, 42 (2), 157–68.

O'Donnell, Roy C. (1974), "Syntactic Differences Between Speech and Writing," *American Speech*, 49 (1/2), 102–10.

Öhman, Arne (1979), "The Orienting Response, Attention and Learning: An Information-Processing Perspective," in *The Orienting Reflex in Humans*, H.D. Kimmel, E.H. van Olst, and J.F. Orlebeke, eds. Mouton, 443–71.

Oksenberg, Lois, Lerita Coleman, and Charles F. Cannell (1986), "Interviewers' Voices and Refusal Rates in Telephone Surveys," *Public Opinion Quarterly*, 50 (1), 97–111.

Ordenes, Francisco Villarroel, Dhruv Grewal, Stephan Ludwig, Ko De Ruyter, Dominik Mahr, and Martin Wetzels (2019), "Cutting Through Content Clutter: How Speech and Image Acts Drive Consumer Sharing of Social Media Brand Messages," *Journal of Consumer Research*, 45 (5), 988–1012.

Packwood, William T. (1974), “Loudness as a Variable in Persuasion,” *Journal of Counseling Psychology*, 21 (1), 1–2.

Pennebaker, James, Ryan Boyd, Kayla Jordan, and Kate Blackburn (2015), “*The Development and Psychometric Properties of LIWC2015,”* University of Texas at Austin.

Perrin, Nicole (2021), “US Digital Ad Spending 2021,” eMarketer (April 14), https://content-na1.emarketer.com/us-digital-ad-spending-2021.

Petkov, Christopher, Christoph Kayser, Thomas Steudel, Kevin Whittingstall, Mark Augath, and Nikos K. Logothetis (2008), “A Voice Region in the Monkey Brain,” *Nature Neuroscience*, 11 (3), 367–74.

Petty, Richard E. and John T. Cacioppo (1986), *Communication and Persuasion: Central and Peripheral Routes to Attitude Change*. Springer.

Pollari, Niina (2015), “A Few Tips on Creating a Good Video,” Kickstarter (July 22), https://www.kickstarter.com/blog/a-few-tips-on-creating-a-good-video.

Poole, Millicent E. and T.W. Field (1976), “A Comparison of Oral and Written Code Elaboration,” *Language and Speech*, 19 (4), 305–12.

Rensink, Ronald (2002), “Change Detection,” *Annual Review of Psychology*, 53, 245–77.

Robinson, Christopher, Robert Moore Jr., and Thomas Crook (2018), “Bimodal Presentation Speeds Up Auditory Processing and Slows Down Visual Processing,” *Frontiers in Psychology*, 9, 2454.

Robinson, Christopher and Vladimir Sloutsky (2019), “Two Mechanisms Underlying Auditory Dominance: Overshadowing and Response Competition,” *Journal of Experimental Child Psychology*, 178, 317–40.

Rose, Robert (1980), “Encoding Variability, Levels of Processing, and the Effects of Spacing of Repetitions upon Judgments of Frequency,” *Memory & Cognition*, 8 (1), 84–93.

Rosenbaum, Paul R. (2020), *Design of Observational Studies*, 2nd ed. Springer.

Rutten, Sanne, Roberta Santoro, Alexis Hervais-Adelman, Elia Formisano, and Narly Golestani (2019), “Cortical Encoding of Speech Enhances Task-Relevant Acoustic Information,” *Nature Human Behaviour*, 3 (9), 974–87.

Scharf, Bertram (1998), “Auditory Attention: The Psychoacoustical Approach,” in *Attention*, Harold Pashler, ed. Psychology Press, 75–117.

Schweinberger, Stefan, Hideki Kawahara, Adrian Simpson, Verena Skuk, and Romi Zäske (2014), “Speaker Perception,” *Wiley Interdisciplinary Reviews: Cognitive Science*, 5 (1), 15–25.

Shrout, Patrick E. and Joseph L. Fleiss (1979), “Intraclass Correlations: Uses in Assessing Rater Reliability,” *Psychological Bulletin*, 86 (2), 420–28.

Siemer, Matthias and Ranier Reisenzein (1998), “Effects of Mood on Evaluative Judgments: Influence of Reduced Processing Capacity and Mood Salience,” *Cognition and Emotion*, 12 (6), 783–805.

Simons, Daniel and Daniel Levin (1997), “Change Blindness,” *Trends in Cognitive Sciences*, 1 (7), 261–67.

Simonson, Itamar (2015), “Mission (Largely) Accomplished: What’s Next for Consumer BDT-JDM Researchers?” *Journal of Marketing Behavior*, 1, 9–35.

Smith, Michael and Takeo Kanade (1998), “Video Skimming and Characterization Through the Combination of Image and Language Understanding,” *Proceedings: 1998 IEEE International Workshop on Content-Based Access of Image and Video Database*, 61–70, https://doi.org/10.1109/CAIVD.1998.646034.

Smith, Stephen M. and David Shaffer (1995), “Speed of Speech and Persuasion: Evidence for Multiple Effects,” *Personality and Social Psychology Bulletin*, 21 (10), 1051–60.

Sokolov, Eugene N. (1963), “Higher Nervous Functions: The Orienting Reflex,” *Annual Review of Physiology*, 25 (1), 545–80.

Spencer, Steven J., Mark P. Zanna, and Geoffrey T. Fong (2005), “Establishing a Causal Chain: Why Experiments Are Often More Effective Than Mediational Analyses in Examining Psychological Processes,” *Journal of Personality and Social Psychology*, 89 (6), 845–51.

Statista (2021), “Kickstarter: Amount of Funding Pledged by Project Category 2020,” (accessed May 13, 2021), https://www.statista.com/statistics/222455/amount-of-dollars-pledged-per-category-on-kickstarter/.

Statista (2023), “Number of Digital Video Viewers Worldwide from 2019 to 2023,” (accessed January 25, 2023), https://www.statista.com/statistics/1061017/digital-video-viewers-number-worldwide/.

Think with Google (2019), “How Online Video Empowers People to Take Action,” (December 9), https://www.thinkwithgoogle.com/feature/youtube-strategy-to-drive-action/.

Treisman, Anne M. and Jenefer G. Riley (1969), “Is Selective Attention Selective Perception or Selective Response? A Further Test,” *Journal of Experimental Psychology*, 79 (1, Pt. 1), 27–34.

Unnava, H. Rao and Robert Burnkrant (1991), “The Effect of Repeating Varied and Same Executions on Brand Name Memory,” *Journal of Marketing Research*, 28 (4), 406–16.

Vajjala, Sowmya and Detmar Meurers (2014), “Exploring Measures of ‘Readability’ for Spoken Language: Analyzing Linguistic Features of Subtitles to Identify Age-Specific TV Programs,” *Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR)*. Association for Computational Linguistics, 21–29.

Voices (2018), “2018 Voice Over Trends in Marketing and Advertising,” (accessed December 9, 2019), https://static.voices.com/assets/uploads/client/2018-Voice-Over-Trends-In-Marketing-v2.pdf.

Von Kriegstein, Katharina, David Smith, Roy Patterson, Stefan Kiebel, and Timothy Griffiths (2010), “How the Human Brain Recognizes Speech in the Context of Changing Speakers,” *Journal of Neuroscience*, 30 (2), 629–38.

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

Younkin, Peter and Venkat Kuppuswamy (2018), “The Colorblind Crowd? Founder Race and Performance in Crowdfunding,” *Management Science*, 64 (7), 3269–87.

YouTube Advertising (2019), “How to Make a Video Ad for Your Marketing Strategy,” (accessed December 9, 2019), https://www.youtube.com/ads/resources/how-to-make-a-video-ad-that-fits-your-marketing-strategy/.

Zhu, Rui (Juliet) and Joan Meyers-Levy (2005), “Distinguishing Between the Meanings of Music: When Background Music Affects Product Perceptions,” Journal of Marketing Research, 1 (3), 333–45.