Contrast Is in the Eye of the Beholder: Infelicitous Beat Gesture Increases Cognitive Load During Online Spoken Discourse Comprehension

Laura M. Morett, Jennifer M. Roche, Scott H. Fraundorf, James C. McPartland

Department of Educational Studies in Psychology, Research Methodology, and Counseling, University of Alabama
Department of Speech Pathology and Audiology, Kent State University
Department of Psychology, Learning Research and Development Center, University of Pittsburgh
Department of Child Psychiatry, Yale University

Received 29 September 2019; received in revised form 15 May 2020; accepted 2 September 2020

Abstract

We investigated how two cues to contrast—beat gesture and contrastive pitch accenting—affect comprehenders’ cognitive load during processing of spoken referring expressions. In two visual-world experiments, we orthogonally manipulated the presence of these cues and their felicity, or fit, with the local (sentence-level) referential context in critical referring expressions while comprehenders’ task-evoked pupillary responses (TEPRs) were examined. In Experiment 1, beat gesture and contrastive accenting always matched the referential context of filler referring expressions and were therefore relatively felicitous on the global (experiment) level, whereas in Experiment 2, beat gesture and contrastive accenting never fit the referential context of filler referring expressions and were therefore infelicitous on the global level. The results revealed that both beat gesture and contrastive accenting increased comprehenders’ cognitive load. For beat gesture, this increase in cognitive load was driven by both local and global infelicity. For contrastive accenting, this increase in cognitive load was unaffected when cues were globally felicitous but exacerbated when cues were globally infelicitous. Together, these results suggest that comprehenders’ cognitive resources are taxed by processing infelicitous use of beat gesture and contrastive accenting to convey contrast on both the local and global levels.

Keywords: Beat gesture; Pitch accent; Linguistic contrast; Visual world; Eye-tracking; Pupillometry

Correspondence should be sent to Laura M. Morett, Department of Educational Studies in Psychology, Research Methodology, and Counseling, University of Alabama, Box 870240, Tuscaloosa, AL 35487.
E-mail: lmorett@ua.edu
1. Introduction

In spoken discourse, multiple cues to referential intent are present in both the auditory and visual modalities. Two such cues are *beat gesture*—simple, rhythmic hand movements—and *pitch accent*—metrical intonational features, both of which convey informational prominence (Bolinger, 1958; McNeill, 1992). Considerable evidence supporting constraint-based models of language comprehension suggests that these and other correlated cues can guide online language processing, leading to more rapid reference resolution (see McRae & Matsuki, 2013, for a review). Although it is evident that these cues ultimately benefit comprehension, it is less clear how they affect comprehenders’ online cognitive load (i.e., cognitive resource consumption) during online language processing, particularly with respect to their relation to one another and the linguistic context in which they occur.

Some theoretical accounts suggest that cognitive load is lightened by additional cues to referential intent, particularly when these cues are consistent with information conveyed via speech. For instance, growth point theory (McNeill, 2005) postulates that gesture and speech arise from unitary semantic representations, suggesting that gesture related to speech should lighten cognitive load by distributing information across the visual, verbal, and haptic channels; in a related vein, dual-coding theory (Paivio, 1990) postulates that visual information complementing information conveyed verbally via speech decreases cognitive load. Thus, these theories predict that beat gesture, particularly in combination with appropriate pitch accenting, may decrease comprehenders’ cognitive load when these cues match—that is, are felicitous with—their referential context. By contrast, theories in which demand for cognitive resources compounds with each additional cue (e.g., Just & Carpenter, 1992) predict that beat gesture, particularly in combination with appropriate pitch accenting, may increase comprehenders’ cognitive load when these cues are felicitous with their referential context. A third class of theories, known as predictive coding accounts (Crosse, Butler, & Lalor, 2015; Olasagasti, Bouton, & Giraud, 2015), posits that the (co-)occurrence statistics of sensory stimuli affect perception. Thus, predictive coding accounts predict that participants’ prior experience observing multimodal emphasis cues co-occurring with contrast—and one another—in natural conversation affects their cognitive resource consumption when these factors are manipulated relative to one another in a laboratory context. In the present work, we investigate how pupil size, an implicit online measure of cognitive load, differs during comprehension of spoken referential expressions conveying (non-) contrastive information (un)accompanied by beat gesture and contrastive pitch accenting in contexts in which these cues are used felicitously and infelicitously, adjudicating between these possibilities and their relationship to discursive features.

1.1. Cues to reference resolution

Successful discourse comprehension entails integrating and decoding a variety of audiovisual cues conveyed via the body. In addition to spoken language, these cues
encompass mouth, tongue, and face movements conveying acoustic information visually (i.e., *visemes*) as well as body movements conveying meaning, rhythm, and attention (*gestures*). Perception of such visual cues to speech activates a network of motor areas in the brain that overlap with those activated by auditory speech and are distinct from those activated by nonspeech biological motion, such as walking (Santi, Servos, Vatikiotis-Bateson, Kuratate, & Munhall, 2003). In addition, integrating visual cues to speech with concurrent spoken language recruits multisensory areas of the brain, resulting in enhanced speech perception (Callan et al., 2003). Although audiovisual integration is necessary and often facilitates speech comprehension, it can result in percepts that correspond neither to the auditory nor to the visual input, but to an intermediate percept, which may add to the speaker’s cognitive load. A classic example is the McGurk effect, which occurs when concurrent visual /g/ and acoustic /b/ are perceived as /d/, but not when concurrent visual /b/ and acoustic /g/ are perceived as /gb/ or /bg/ (Olasagasti et al., 2015). Thus, understanding how audiovisual cues are integrated and how this integration affects the listener’s cognitive load is crucial to understanding how discourse is comprehended.

Another crucial aspect of successful discourse comprehension is establishing relations between referring expressions and their referents. In a context in which a red and blue ball are present, upon hearing *Bring me a ball/Now bring me another ball*, the listener can infer that the talker doesn’t care which color ball is retrieved each time as long as each ball is retrieved in succession. By comparison, within the same context, upon hearing *Bring me the red ball/Now bring me the blue ball*, the listener can infer that the talker wants him to retrieve the red ball and then the blue ball in that order.

The distinction between the two referents (red ball and blue ball) highlighted in each referring expression (*the red ball, the blue ball*) is known as *contrast*, which refers to a contradiction between two discursive themes (in this case, color; Myhill & Xing, 1996). Although modifiers, such as color adjectives, are commonly used to convey contrast, modifier subclasses vary in contextual sensitivity. For instance, scalar and material modifiers, such as *large* and *wooden*, are often used specifically in contrastive contexts (e.g., contexts containing two balls differing in size or composition; see Sedivy, Tanenhaus, Chambers, & Carlson, 1999). However, color is often used attributively even in non-contrastive contexts (e.g., contexts containing a single ball, which happens to be red). In light of the unreliability of color modifiers as indicators of contrast, it may behoove comprehenders to take other cues into account when determining whether two referents contrast.

### 1.1.1. Prosodic prominence

One such additional cue that can be used to convey contrast in discourse is prosodic prominence. Prominence can be signaled by several acoustic features, including variation in duration, intensity (loudness), and fundamental frequency ($F_0$); in some cases, variation in prominence stems from the acoustic realization of *pitch accents*, phonological constructs that may be placed on specific words¹ (for a review, see Ladd, 2008). Prominence can vary for multiple reasons (Watson, 2010), but one generalization is that information that is contrastive is often acoustically more prominent (Selkirk, 2002). For example, in the discourse, *What type of pet does Maya have?*/Maya has a black cat, *black* is a new
color modifier that doesn’t contrast with any proximal color modifiers. Thus, it would likely be less prominent than in the discourse, *Did you say Maya has a brown cat?* / *No, I said Maya has a BLACK cat*, in which *black* contrasts with the previously mentioned color modifier *brown*. Although it has been debated whether this variation reflects a qualitative difference between contrastive and non-contrastive pitch accents (e.g., Pierrehumbert & Hirschberg, 1990; Selkirk, 2002) or a quantitative difference (e.g., Dilley, 2010; Ladd & Schepman, 2003), there is ample evidence that intonation can ultimately convey contrast to comprehenders (e.g., Fraundorf, Watson, & Benjamin, 2010; Ito & Speer, 2008; Watson, Tanenhaus, & Gunlogson, 2008). Thus, in the current work, we investigate how differences in contrastive versus non-contrastive intonation affect cognitive load without making assumptions about the underlying linguistic representations that give rise to this variation in pitch accent.

1.1.2. Beat gesture and its relation to pitch accenting

Another cue that can be used to convey contrast in discourse is co-speech gesture—meaningful movements of the hands or other body parts. According to a widely used gesture taxonomy (McNeill, 1992, 2005), all co-speech gestures can be classified as at least one of four types: *iconic gestures*, which convey actions or affordances of concrete referents; *metaphorical gestures*, which convey actions or affordances of the sources of metaphorical referents; *deictic gestures*, which index referents via pointing; or *beat gestures*, which convey emphasis via flicks beating time along with the rhythm of speech. Although several of these types of gesture can be used to convey contrast, in this work, we focus on beat gesture due to its close relationship to pitch accent—including contrastive accenting—in both timing and function.

With respect to timing, beat gesture tends to co-occur with pitch accented syllables (Yasinnik, Renwick, & Shattuck-Hufnagel, 2004), and its *apices*, or points of furthest extension, are tightly temporally aligned with $F_0$ peaks in speech (Leonard & Cummins, 2011). Similar alignment is observed with deictic gesture that functions similar to beat gesture in a context featuring contrastive focus (Esteve-Gibert & Prieto, 2013). Moreover, utterances in which beat gesture and pitch accent align on the same word are rated as more natural than utterances in which beat gesture and pitch accent occur on different words (Krahmer, Ruttkay, Swerts, & Wesselink, 2002). In addition, manipulating syllable position, contrastive stress, and auditory feedback through a delay disrupts the timing of deictic gesture that functions similar to beat gesture (Rusiewicz, Shaiman, Iverson, & Szuminsky, 2013, 2014). Thus, the dynamical system responsible for motor coordination has been postulated as the mechanism through which the close temporal relationship between discrete gesture (i.e., beat and similar deictic gesture) and pitch accent arises (Port, Cummins, & Gasser, 1995; Rusiewicz, 2011).

Although these findings suggest beat gesture can align with acoustic peaks, acoustic peaks are not necessarily cognitively salient (Barnes, Veilleux, Brugos, & Shattuck-Hufnagel, 2012; Dilley & Heffner, 2013; House, 1992). Nevertheless, other evidence suggests a link in perception as well: Beat gesture increases the perceived prominence of pitch accented words and decreases that of unaccented words (Ambrazaitis & House, 2017;
Krahmer & Swerts, 2007). Moreover, words accompanied by beat gesture are produced with higher vowel formants and more emphasis and are perceived as more prominent than words unaccompanied by beat gestures (Krahmer & Swerts, 2007). These findings suggest that the combination of beat gesture and pitch accent is particularly effective at conveying emphasis. With respect to contrast in particular, emerging evidence indicates that, when beat gesture and contrastive pitch accent both co-occur with contrasting information in spoken discourse such that these cues are locally felicitous with contrast, comprehenders’ subsequent memory for this contrasting information is enhanced (Llanes-Coromina, Vilà-Giménez, Kushch, Borràs-Comes, & Prieto, 2018; Morett & Fraundorf, 2019). Thus, co-occurrence of beat gesture with contrastive accenting—and contrastive information in discourse—may be critical to its interpretation as a cue to contrast.

1.2. The cognitive load of processing linguistic cues

Do these additional cues to reference resolution reduce or increase the cognitive load of online language processing? Some theories propose that the combination of visual beat gesture and auditory contrastive accenting should be especially efficacious in reducing comprehenders’ cognitive load during language processing. For instance, dual-coding theory (Paivio, 1990) postulates that when complementary information pertinent to a concept is available in both the visual and verbal (auditory) modalities, it is processed via both visual and verbal working memory, leading to superior comprehension, memory, and learning of that concept. Evidence from dual-task paradigms supports these predictions: Instructional materials presented both visually and verbally tax cognitive load less than instructional materials presented only visually (Brünken, Plass, & Leutner, 2004; Brünken, Steinbacher, Plass, & Leutner, 2002), indicating that multimodal processing demands fewer cognitive resources than unimodal processing of the same information. A seeming paradox of dual-coding theory is that the amount of information distributed across both modalities exceeds comprehenders’ processing capacity yet is more likely to be learned than the same amount of information presented within a single modality (Mayer & Moreno, 2003; Moreno, Mayer, Spires, & Lester, 2001; Moreno et al., 2001; Mousavi, Low, & Sweller, 1995). This outcome is due to greater efficiency of processing complementary information pertinent to a concept that is presented in two modalities than the same information when it is presented in a single modality, which may be explained by cognitive load.

On the other hand, additive theories of cue integration in which further cognitive processing resources are required for each additional cue or constraint on linguistic processing (e.g., CC-READER; Just & Carpenter, 1992) make the converse prediction that multimodal cues should tax comprehenders’ cognitive load during language processing, particularly when cognitive resources are limited. In a unimodal context, there is evidence that animacy of the head noun fails to facilitate reduced relative clause processing during reading (Ferreira & Clifton, 1986) and that this effect is due to low verbal working memory capacity (Just & Carpenter, 1992). We are unaware of any research to date examining the predictions of additive theories of cue integration with cues in different modalities,
however; therefore, at present, it is unclear whether they accurately account for multi-modal cue integration.

A third class of theories, known as predictive coding frameworks (Crosse et al., 2015; Olasagasti et al., 2015), postulates that comprehenders draw upon the observed hierarchical structure and statistics of discourse to generate top-down predictions of the causes of cues such as contrastive accenting and beat gesture that must be reconciled with prediction errors concerning how these cues are used in a given case. Thus, if contrastive accenting *always* co-occurs with contrast whereas beat gesture *sometimes* co-occurs with contrast (and, therefore, contrastive accenting), contrastive accenting that occurs in the absence of contrast should increase comprehenders’ cognitive load significantly, whereas beat gesture that occurs in the absence of contrast (and contrastive accenting) should increase comprehenders’ cognitive load moderately in accordance with how infrequently it has previously occurred in a similar context.

1.2.1. The cognitive load of gesture and pitch accenting

Previous research provides evidence that both producing and viewing gestures of various types may lighten the respective cognitive loads of talkers and comprehenders during language processing. With respect to production, gesturing increases during cognitively complex tasks (Melinger & Kita, 2007) and improves secondary task performance in dual-task paradigms (Goldin-Meadow, Nusbaum, Kelly, & Wagner, 2001; Melinger & Kita, 2007; Ping & Goldin-Meadow, 2010; Wagner, Nusbaum, & Goldin-Meadow, 2004). Moreover, individuals with low working memory capacity are more likely than individuals with high working memory capacity to gesture during tasks with high working memory demand (Gillespie, James, Federmeier, & Watson, 2014; Marstaller & Burianová, 2013). However, some evidence suggests that mimicking gestures redundant with speech may hinder learning (Byrd, McNeil, D’Mello, & Cook, 2014). Nevertheless, taken together, these findings suggest that spontaneous gesture produced concurrently with speech may serve as a means of cognitive offloading for talkers (Goldin-Meadow, 2003).

With respect to viewing, seeing talkers produce relevant gesture scaffolds second-language vocabulary acquisition in comprehenders (Lazaraton, 2004; McCafferty, 2002; Morett & Chang, 2015). In a related vein, viewing beat gesture enhances memory for words from one’s native language (Igualada, Esteve-Gibert, & Prieto, 2017; So, Sim Chen-Hui, & Low Wei-Shan, 2012). Moreover, viewing beat gesture enhances memory for words and discourse in both the first and second languages, with some work showing a beneficial effect of beat gesture alone (Austin & Sweller, 2014; Vilà-Giménez, Igualada, & Prieto, 2019) and other work showing a beneficial effect only for beat gesture occurring concurrently with pitch accenting (Kushch, Igualada, & Prieto, 2018; Kushch & Prieto, 2016; Llanes-Coromina et al., 2018; Morett & Fraundorf, 2019). Mixed results have been observed with respect to the effect of gesture on learning complex mathematical concepts such as equivalence, symmetry, and slope, however, with some work demonstrating a beneficial effect of viewing gesture (Alibali & Nathan, 2007; Valenzeno, Alibali, & Klatzky, 2003) and other work demonstrating that viewing gesture hinders learning (Byrd et al., 2014; Yeo, Ledesma, Nathan, Alibali, & Church, 2017).
Nevertheless, the extant literature broadly suggests that viewing co-speech gesture related to information conveyed via spoken discourse may decrease comprehenders’ cognitive load.

To date, however, we are not aware of any research examining the predictions of dual coding theory, additive theories of cue integration, or predictive coding frameworks concerning the effects of multimodal cues on cognitive load during language processing using online measures. The current research fills this lacuna, offering insight into which of these classes of theories most accurately characterizes the impact of beat gesture on comprehenders’ cognitive load, as well as whether growth point theory accurately characterizes the impact of beat gesture’s congruence with referential contrast on comprehenders’ cognitive load during spoken discourse processing. It does so using pupillometry, a powerful method that we now discuss.

1.2.2. Pupillometry and cognitive load in language processing

Research examining the task-evoked pupillary response (TEPR) provides evidence that the relationship between pitch accenting and referent status affects comprehenders’ cognitive load during spoken discourse processing. Generally, smaller pupillary responses are observed when language processing is facilitated by intrinsic characteristics of the linguistic input such as speech quality (Koch & Janse, 2016; Kuchinsky et al., 2013; McGarrigle, Dawes, Stewart, Kuchinsky, & Munro, 2017; Tamási, McKeans, Gafos, Fritzscbe, & Höhle, 2017; Wagner, Toffanin, & Başkent, 2016; Winn, Edwards, & Litovský, 2015; Zekveld, Kramer, & Festen, 2010), lexical characteristics (Chapman & Hallowell, 2015; Frank & Thompson, 2012; Geller, Landrigan, & Mirman, 2019; Geller, Still, & Morris, 2016; Guasch, Ferre, & Haro, 2017; Haro, Guasch, Vallès, & Ferré, 2017; Hyönä, Tommola, & Alaja, 1995; Kuchinke, Võ, Hofmann, & Jacobs, 2007; Ledoux et al., 2016; Loo, van Rij, Järvikivi, & Baayen, 2016; Papesh & Goldinger, 2012), lack of syntactic ambiguity (Ben-Nun, 1986; Just & Carpenter, 1993; Nikuni, Yasunaga, Iwasaki, & Muramoto, 2015; Sauppe, 2017; Schluoroff, 1982; Sevilla, Maldonado, & Shalóm, 2014), and discursive implications (Demberg & Sayeed, 2016; Tromp, Hagoort, & Meyer, 2016). Conversely, larger pupillary responses are observed when language must be comprehended while simultaneously performing other tasks (Causse, Peysakhovich, & Fabre, 2016; Demberg & Sayeed, 2016; Koelewin, de Kluiver, Shinn-Cunningham, Zekveld, & Kramer, 2015; Koelewin, Shinn-Cunningham, Zekveld, & Kramer, 2014; Kramer et al., 2013) as well when language comprehension is difficult due to conditions such as aphasia and cognitive aging (Chapman & Hallowell, 2015; Hochmann & Papeo, 2014; Koch & Janse, 2016; Piquado, Isaacowitz, & Wingfield, 2010; Schmidtke, 2014; Wendt, Dau, & Hjortkjaer, 2016; see Schmidtke, 2018, for a review), indicating that task-external factors also affect pupillary responses during language comprehension.

This relationship extends to the prosodic cues of interest in the present study. For instance, comprehenders’ pupil size is smaller when speech prosody is consistent (vs. inconsistent) with syntactic structure (Engelhardt, Ferreira, & Patsenko, 2010; Scheepers, Mohr, Fischer, & Roberts, 2013). Most critically for our purposes, when a correction is made involving a referential contrast such as Did you say Maya has a brown cat?/No, I
said Maya has a BLACK cat, comprehenders’ pupil size is smaller when the contrasting information has contrastive rather than non-contrastive (presentational) accenting. For non-contrastive information such as What type of pet does Maya have?/Maya has a black cat, however, pupil size does not differ as a function of pitch accent type (Zellin, Pannekamp, Toepel, & van der Meer, 2011). This finding suggests that when contrastive pitch accenting conveys referential contrast, it facilitates language comprehension and thereby decreases comprehenders’ cognitive load.

However, much of the extant work using pupillometry to examine cognitive load during language processing has considered the effect of a single cue—such as pitch accenting—in isolation. Here, we consider the effects of multiple cues in different modalities (i.e., beat gesture, pitch accenting) on comprehenders’ cognitive load by varying the presence of beat gesture and contrastive accenting orthogonally relative to one another, allowing us to assess how these cues affect cognitive load both individually and in interaction.

1.2.3. Global felicity and cognitive load

A second goal of the present study was to examine how processing of these cues—and, in particular, their effect on cognitive load—is influenced by their global felicity. As noted above, predictive coding frameworks (Crosse et al., 2015; Olasagasti et al., 2015) propose that the effects of these cues depend on their congruency within the broader context of comprehenders’ previous experience with similar discourse.

Indeed, emerging evidence now suggests that interpretation of referential cues may depend on their broader use within a given communicative context. In a pioneering study, Grodner and Sedivy (2011) examined how providing a meta-linguistic explanation (that the talker had an impairment negating pragmatic competence) as well as evidence of the talker’s language use patterns (over-informative use of scalar adjectives in filler trials) affected interpretation of contrastive referring expressions containing scalar adjectives. This work revealed that, on critical trials, comprehenders were less likely to look at the target referent given this cover story than when no information about the talker was provided and scalar adjectives were used informatively in filler trials. More recent work using a similar paradigm indicates that, in sufficiently large quantities, evidence of pragmatically inappropriate use of scalar adjectives influences referring expression interpretation even in the absence of a meta-linguistic explanation (Ryskin, Kurumada, & Brown-Schmidt, 2019). These findings are consistent with our other work, which indicates that interpretation of both beat gesture and contrastive accenting as cues to contrast depends not only on their local (i.e., sentence-level) felicity with contrastive information in target discourse, but also on their global (i.e., experiment-level) felicity with contrastive information in filler discourse (Morett et al., in press).

Although these findings suggest that whether various cues occur felicitously within a given experimental context may affect eventual reference resolution, it is currently unclear how such variation in cue felicity affects comprehenders’ cognitive load during online language comprehension. Here, we consider how the co-occurrence of beat gesture and contrastive accenting with contrastive information affects comprehenders’ cognitive load by varying the match between these cues and contrast in critical and filler referring expressions.
1.3. Present study

The present work is the first to investigate how beat gesture and contrastive accenting affect comprehenders’ cognitive load during online spoken discourse processing. In particular, we examine the effects of these cues on cognitive load during interpretation of pairs of referential expressions varying in contrast.

We use the TEPR to assess the effects of beat gesture and contrastive accenting on comprehenders’ cognitive load during referential expression processing. The TEPR is an ideal measure for investigating this question because, when controlling other factors known to affect pupil size such as ambient luminance and affect, the TEPR provides an online and implicit measure of cognitive processing effort (Beatty, 1982; Brünken, Plass, & Leutner, 2003; Hess & Polt, 1964; Kahneman & Beatty, 1966; Laeng, Sirois, & Gredebäck, 2012; Mathôt, 2018). As a test of comprehenders’ ability to process these referential expressions, we adapted the visual world paradigm, in which spoken language is presented with a concurrent display containing visual representations of referents (Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995). Our variant of this paradigm introduced a video of a talker who sometimes produced beat gesture when referring to target items (see also Silverman, Bennetto, Campana, & Tanenhaus, 2010). The TEPR has been successfully utilized in conjunction with the visual world paradigm in several previous studies (Koch & Janse, 2016; Kuchinsky et al., 2013; Ledoux et al., 2016; Wagner et al., 2016), demonstrating that it can reveal systematic variation in comprehenders’ online cognitive load resulting from manipulations of key features of spoken discourse.

We consider three ways—which are not mutually exclusive—in which beat gesture and contrastive accenting might affect comprehenders’ cognitive load. First, the presence of these cues might affect cognitive load regardless of referential context based on, for instance, the visual complexity of beat gesture or the acoustic complexity of contrastive accenting. In this case, the presence of beat gesture and contrastive accenting should elicit effects that do not vary based on their felicity with their referential contexts. Second, these cues’ impact on cognitive load might be affected by their local felicity with referential context on the sentence level. In this case, cognitive load may be lower when these cues co-occur with contrasting color words, as in Click on the purple star/Now click on the red star, supporting reference resolution, but higher when these cues co-occur with color words in sentences in which both color and shape differ, as in Click on the purple heart/Now click on the red star, hindering reference resolution. Third, these cues’ impact on cognitive load might be affected by their global felicity with referential context on the experiment level. In this case, cognitive load may be lower in an experimental context in which those cues are often felicitous with each other and with contrast, but higher within an experimental context in which these cues are often infelicitous with each other and with contrast. In addition to these possibilities, the effects of beat gesture and contrastive accenting might differ in direction and/or magnitude, and their effects might interact.

To distinguish between these possibilities, we conducted two experiments in which we orthogonally manipulated the presence of beat gesture, contrastive accenting, and their fit
to their local and global referential contexts. In both experiments, the presence of beat gesture and contrastive accenting were independently manipulated in conjunction with words that specifically contrasted or did not contrast in critical items, allowing us to investigate how each of these cues’ local felicity with contrast affected reference resolution. In Experiment 1, filler items contained beat gesture and contrastive accenting on contrasting words, such that these cues were globally felicitous; this design should encourage interpretation of beat gesture and contrastive accenting as cues to reference resolution in critical items. In Experiment 2, filler items contained beat gesture and contrastive accenting that were orthogonally manipulated on non-contrasting words, such that these cues were globally infelicitous; this design should discourage interpretation of beat gesture and contrastive accenting as cues to reference resolution in critical items because they are often uninformative with respect to target referent identity. Although several cue and felicity combinations entailed by these manipulations are typically unattested within natural spoken discourse, they are nevertheless crucial to tease apart the effects of cues themselves from their felicity with their local and global referential contexts on cognitive load during spoken discourse comprehension. Thus, together, these experiments allow us to assess the degree to which beat gesture and contrastive accenting’s effects on comprehenders’ cognitive load are driven by local vs. global felicity of these cues.

2. Experiment 1: Felicitous context

In Experiment 1, we examined the effects of beat gesture and contrastive accenting on cognitive load during spoken discourse comprehension within a context in which these cues were globally felicitous. To examine the independent and combined effects of beat gesture and contrastive accenting on cognitive load within this context, these cues were manipulated orthogonally on color words in critical sentences, whereas contingencies between these cues and contrasting shape words were maintained in filler sentences. Thus, Experiment 1 tested whether beat gesture and contrastive accenting, which convey contrast beyond lexical information, facilitate interpretation of color-contrast critical referring expressions in a global experimental context in which these cues occurred felicitously in filler items.

2.1. Methods

2.1.1. Participants

Forty adult monolingual native English speakers (age range: 18–35 years; 29 women, 11 men) participated in Experiment 1 and a separate, unrelated electrophysiological study in return for $25 USD. All participants were recruited from the community surrounding a large private research university in the Northeast United States via electronic and hard copy announcements. All participants reported having normal hearing, normal or corrected-to-normal vision, and no color blindness.
2.1.2. Design

Contrast type, beat gesture, and pitch accenting were manipulated orthogonally in critical referring expressions, resulting in a 2 (color-contrast vs. color-+shape-difference) × 2 (contrastive vs. non-contrastive accenting) × 2 (beat vs. no beat) design. Eight lists counterbalanced by these independent variables (Lists 1.1–2.4) were used in both practice and experimental trials. Because beat gesture and pitch accenting were manipulated on color words, their presence was locally felicitous in the color-contrast condition and locally infelicitous in the color-+shape-difference condition, whereas the opposite was true for their absence. Fig. 1 displays the sentences and accompanying visual displays used in the first critical and filler trials of Lists 1.1–1.4. (See Table S1 in Supplementary Material for additional information concerning trials presented in each list.)

2.1.3. Materials

2.1.3.1. Objects: A total of 64 objects (representing all possible combinations of the eight colors and eight shapes listed in Table 1) were created for use in arrays accompanying audio and video stimuli during practice and experimental trials. Objects presented during critical referring expressions consisted of four types (see Fig. 1 for example): The context object was the referent of the context referring expression (e.g., blue star), the target object was the referent of the critical referring expression (e.g., white star [color-contrast] or white square [color-+shape-difference]), the competitor object was an object with the same color but an alternate shape relative to the target object (e.g., white square [color-contrast] or white star [color-+shape-difference]), and the distractor object was an object with the remaining combination of colors and shapes from other objects in the array (e.g., blue square). Critically, the presence of the competitor object created a
temporary referential ambiguity (*Now click on the white...*) between the target and competitor objects.

Across all trials, objects corresponding to all four possible continuations (color-contrast, color- + shape-difference, shape-contrast, neither-difference) were present in the array, and objects corresponding to all four possible continuations were equally likely to follow the context object. Each of the 64 objects was assigned as the context object, target object, competitor object, and distractor object in two to three trials in a variety of conditions in order to avoid contingencies to the greatest extent possible; however, a perfectly equal number of assignments for each individual object was not possible because the number of trials was specified based on the independent variables (contrast type, pitch accent, and beat gesture), and we prioritized counterbalancing of those variables.

During all trials, objects appeared in one of four locations arranged in a square configuration surrounding centrally presented, circularly framed videos (see Fig. 1). For each trial, positioning of objects was counterbalanced across participants such that context, target, competitor, and distractor objects were equally likely to appear in each position (see Tables S4 and S5).

2.1.3.2. Audio recordings: All sentences audio recorded for use in the experiment were produced by an adult female monolingual speaker of Standard American English. Sentences were recorded using professional equipment in a sound-shielded room, prior to video recording, to ensure they were of the highest possible quality.

Each sentence referred to a specific object from an array (see Fig. 1 for example arrays and 7–10d for example sentences). Each trial consisted of a pair of sentences, the first referring to a context object (context sentence; 7, 9) and the second referring to a critical object (continuation sentence; 8a–d, 10a–d; see Appendix A, publicly available at https://osf.io/fy6wp/, for complete list of sentences used in Experiment 1). Within each stimulus list used in the experiment (Lists 1.1–1.4 and 2.1–2.4), referring expressions within context and continuation sentences were counterbalanced, such that each critical sentence appeared in trials with all possible contrast types. (See Table S2 in the Supplementary Material for the number of sentences constructed with each pitch accent within each list set.)

<table>
<thead>
<tr>
<th>Colors</th>
<th>Shapes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Red</td>
<td>Circle</td>
</tr>
<tr>
<td>Blue</td>
<td>Square</td>
</tr>
<tr>
<td>Green</td>
<td>Triangle</td>
</tr>
<tr>
<td>Yellow</td>
<td>Rectangle</td>
</tr>
<tr>
<td>Orange</td>
<td>Diamond</td>
</tr>
<tr>
<td>Purple</td>
<td>Oval</td>
</tr>
<tr>
<td>White</td>
<td>Star</td>
</tr>
<tr>
<td>Black</td>
<td>Heart</td>
</tr>
</tbody>
</table>
During audio recording, the talker referred to objects from the same array in both context and continuation sentences, the latter of which could have any combination of colors and shapes represented in the array (see Table 1).

For each critical continuation sentence, we created one recording in the contrastive accenting condition and one in the non-contrastive accenting condition. In the contrastive accenting condition, the talker was instructed to emphasize the color adjective and produced the context sentence followed by the corresponding color-contrast continuation sentence; in the non-contrastive accenting condition, the talker was instructed not to emphasize any particular word and produced the context sentence followed by the corresponding color- and shape-difference continuation sentence (see Appendix B, publicly available with other appendices at https://osf.io/fy6wp/, for audio recording list). Acoustic analyses, reported below, confirmed that, as expected, the contrastive accenting condition resulted in greater acoustic prominence (Selkirk, 2002), supporting the validity of our manipulation.

The entire list of sentences was recorded twice in two separate sessions. One recording of each critical sentence with non-contrastive accenting was selected for use of its color word (non-contrastive accenting word sentence), and the other recording was selected for use of its other components (carrier sentence). In addition, one recording of each critical sentence with contrastive accenting was selected for use of its color word (contrastive accenting word sentence). Finally, one recording of each filler sentence with non-contrastive and contrastive accenting was selected for wholesale use. When possible, recordings were selected on the basis of their quality and the prototypicality of their pitch accents. Audio files of final sentences were duplicated as needed and used in all trials in which they were required based on stimulus lists and the experimental design.

2.1.3.3. Critical sentences: In critical experimental trials, the target object took one of two forms: Relative to the preceding (context) object, half differed in color but not shape (color-contrast referring expressions; 2a, 4a), and half differed in both color and shape (color- + shape-difference expressions; 2b, 4b). In critical sentences in these trials, half of the color words received a contrastive pitch accent, whereas the other half of the color
words received a non-contrastive (presentational) accent. All shape words in these sentences were naturally de-accented. Thus, contrastive accenting of color words was locally felicitous in color-contrast but not color-+shape-difference referring expressions, whereas non-contrastive accenting of color words was locally felicitous in color-+shape-difference but not color-contrast referring expressions.

Using Praat (Boersma & Weenink, 2016), text grids for non-contrastive and contrastive accenting word sentences and carrier sentences were created in which initial portions (Now click on the…), color words (purple), stressed syllables of multisyllabic color words (purp-), and shape words (triangle) were annotated. These text grids were subsequently used to batch extract these components as individual files. For color words and stressed syllables of multisyllabic color words in sentences with non-contrastive and contrastive accenting, we measured intensity, duration, maximum $F_0$ (fundamental frequency), differences between maximum and minimum $F_0$, and mean $F_0$.

To verify that color words indeed differed between the contrastive and non-contrastive accenting conditions, we compared their acoustic properties in a series of paired-sample $t$ tests. Because it is argued that pitch accenting is realized on the syllable carrying the primary stress of a word (Ladd, 2008), analyses were conducted on both the stressed syllable of multisyllabic color words as well as entire color words. Table 2 presents means and standard deviations of acoustic measures. Both stressed syllables and color words with contrastive and non-contrastive accenting differed reliably on all measures except for $F_0$ difference.

To eliminate any spurious differences in intensity across individual recordings, intensity was normalized to the average of each component (initial, color word, shape word) for each critical sentence type (word, carrier). Finally, color words from non-contrastive and contrastive accenting word sentences were spliced between initial portions and shape words of carrier sentences. This ensured that the critical word was the only part of critical sentences that differed acoustically or phonologically between the two types of referring expressions. To ensure that splicing of critical sentences was not noticeable, a subset of critical and unspliced sentences was played to a pilot participant, who was unable to reliably identify which sentences were spliced.

2.1.3.4. Filler sentences: Filler trials were included to ensure that all types of objects sometimes served as the target in the continuation referring expression, thus preventing participants from predicting referents based on within-experiment contingencies. Thus, relative to the preceding context referring expression, half of filler referring expressions differed in shape but not color (shape-contrast; 8c, 8d), and half differed in neither shape nor color (neither-difference; 8c, 8d). Shape words always had contrastive accenting in shape-contrast referring expressions and non-contrastive accenting in neither-difference referring expressions, such that pitch accenting was globally felicitous in filler trials. (See Table S3 in the Supplementary Material for a summary of filler item attributes.)

Filler sentences were annotated, segmented, and analyzed for acoustic properties similar to critical sentences. For fillers, we expected the primary accent to fall on the shape word (e.g., triangle) because the fillers either involved a shape contrast or no contrast, in
which the pitch accent would be expected to fall on the head shape noun (e.g., Selkirk, 1995). Thus, acoustic analyses were based on shape words (e.g., \textit{triangle}) and stressed syllables of multisyllabic shape words (e.g., \textit{tri-}). To verify that the contrastive and non-contrastive accenting conditions differed acoustically, we compared them using a series of paired-sample \textit{t} tests. Table 2 presents means and standard deviations of acoustic measures. Stressed syllables with contrastive and non-contrastive accenting differed reliably on all measures except maximum \( F_0 \) and \( F_0 \) difference, whereas shape words with contrastive and non-contrastive pitch accenting differed reliably on all measures except intensity.

<table>
<thead>
<tr>
<th>Measure</th>
<th>Non-contrastive</th>
<th>Contrastive</th>
<th>( df )</th>
<th>( t )</th>
<th>( p )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stressed syllable</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Critical sentence (color)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intensity (dB)</td>
<td>35.33 (1.50)</td>
<td>38.30 (1.66)</td>
<td>58</td>
<td>(-7.27)</td>
<td>(&lt;.001^{***})</td>
</tr>
<tr>
<td>Duration (s)</td>
<td>0.15 (0.02)</td>
<td>0.18 (0.03)</td>
<td>58</td>
<td>(-4.53)</td>
<td>(&lt;.001^{***})</td>
</tr>
<tr>
<td>Maximum ( F_0 ) (Hz)</td>
<td>224.67 (26.17)</td>
<td>330.34 (53.79)</td>
<td>58</td>
<td>(-9.68)</td>
<td>(&lt;.001^{***})</td>
</tr>
<tr>
<td>( F_0 ) difference (Hz)</td>
<td>28.41 (17.58)</td>
<td>96.46 (57.35)</td>
<td>58</td>
<td>(-6.21)</td>
<td>(&lt;.001^{***})</td>
</tr>
<tr>
<td>Mean ( F_0 ) (Hz)</td>
<td>208.53 (21.20)</td>
<td>275.08 (37.26)</td>
<td>58</td>
<td>(-8.50)</td>
<td>(&lt;.001^{***})</td>
</tr>
<tr>
<td>Entire word</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intensity (dB)</td>
<td>35.60 (1.82)</td>
<td>38.04 (1.90)</td>
<td>158</td>
<td>(-8.31)</td>
<td>(&lt;.001^{***})</td>
</tr>
<tr>
<td>Duration (s)</td>
<td>0.26 (0.05)</td>
<td>0.29 (0.06)</td>
<td>158</td>
<td>(-4.08)</td>
<td>(&lt;.001^{***})</td>
</tr>
<tr>
<td>Maximum ( F_0 ) (Hz)</td>
<td>315.45 (141.38)</td>
<td>357.81 (73.33)</td>
<td>158</td>
<td>(-2.38)</td>
<td>.02</td>
</tr>
<tr>
<td>( F_0 ) difference (Hz)</td>
<td>122.35 (141.30)</td>
<td>146.52 (75.74)</td>
<td>158</td>
<td>(-1.35)</td>
<td>.18</td>
</tr>
<tr>
<td>Mean ( F_0 ) (Hz)</td>
<td>217.06 (21.61)</td>
<td>280.89 (21.34)</td>
<td>158</td>
<td>(-18.8)</td>
<td>(&lt;.001^{***})</td>
</tr>
<tr>
<td>Filler sentence (shape)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stressed syllable</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intensity (dB)</td>
<td>37.37 (2.00)</td>
<td>39.16 (1.85)</td>
<td>58</td>
<td>(-3.62)</td>
<td>(&lt;.001^{***})</td>
</tr>
<tr>
<td>Duration (s)</td>
<td>0.20 (0.02)</td>
<td>0.24 (0.05)</td>
<td>58</td>
<td>(-4.83)</td>
<td>(&lt;.001^{***})</td>
</tr>
<tr>
<td>Maximum ( F_0 ) (Hz)</td>
<td>370.15 (170.52)</td>
<td>361.29 (117.67)</td>
<td>58</td>
<td>0.23</td>
<td>.82</td>
</tr>
<tr>
<td>( F_0 ) difference (Hz)</td>
<td>177.81 (176.45)</td>
<td>124.05 (121.38)</td>
<td>58</td>
<td>1.37</td>
<td>.18</td>
</tr>
<tr>
<td>Mean ( F_0 ) (Hz)</td>
<td>235.48 (40.05)</td>
<td>278.40 (31.88)</td>
<td>58</td>
<td>(-4.59)</td>
<td>(&lt;.001^{***})</td>
</tr>
<tr>
<td>Entire word</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intensity (dB)</td>
<td>37.24 (1.98)</td>
<td>36.97 (1.81)</td>
<td>158</td>
<td>0.9</td>
<td>.37</td>
</tr>
<tr>
<td>Duration (s)</td>
<td>0.34 (0.07)</td>
<td>0.52 (0.09)</td>
<td>158</td>
<td>(-13.9)</td>
<td>(&lt;.001^{***})</td>
</tr>
<tr>
<td>Maximum ( F_0 ) (Hz)</td>
<td>383.11 (171.64)</td>
<td>435.82 (137.17)</td>
<td>158</td>
<td>(-2.15)</td>
<td>.03</td>
</tr>
<tr>
<td>( F_0 ) difference (Hz)</td>
<td>208.57 (172.45)</td>
<td>266.63 (147.95)</td>
<td>158</td>
<td>(-2.29)</td>
<td>.02</td>
</tr>
<tr>
<td>Mean ( F_0 ) (Hz)</td>
<td>226.09 (47.56)</td>
<td>261.82 (39.81)</td>
<td>158</td>
<td>(-5.15)</td>
<td>(&lt;.001^{***})</td>
</tr>
</tbody>
</table>

\*\( p < .05; \quad \)**\( p < .001. \)

Again, to eliminate any differences in intensity across individual recordings, intensity was amplified and normalized to the average for filler sentences with each type of accenting (non-contrastive, contrastive). Unlike critical sentences, filler sentences were not segmented or spliced. Pupilometric and mouse click responses to filler sentences were not analyzed because we were interested in examining how orthogonally manipulating beat
gesture and contrastive accenting on the color word affected interpretation of referring expressions either contrasting in color or differing in both color and shape.

2.1.3.5. Video recordings: Videos to accompany all sentences presented in the experiment were recorded separately from audio recordings. Videos featured a Caucasian adult female (henceforth, the *talker*) and were framed on her torso with her head excluded to ensure that facial cues could not affect sentence interpretation. To demonstrate to the talker how beat gesture should be produced during video recording of sentence production, the first author (L.M.M.) modeled one of the most common beat gestures that talkers produce in natural conversation—a single downward flip of the hand with the palm open upward (McNeill, 1992). During video recording, the talker listened to audio recordings of sentences and subsequently repeated them while either producing a beat gesture similar to the one modeled in conjunction with the color or shape word or keeping the hands still (see Table S6 in Supplementary Material for the number of videos constructed for each set of lists). In light of work demonstrating that temporal synchrony with speech prosody defines beat gesture (see note 2), such that any movement temporally synchronous with speech prosody would be interpreted as beat gesture, we decided on a no-movement condition as the most appropriate comparison condition for the beat gesture condition. Thus, the talker was instructed to remain still while producing context sentences (7, 9).

Prior to postproduction, we reviewed the videos and confirmed that the talker produced the correct pitch accent on the correct word and either produced beat gesture or kept her hands still in conjunction with this word, as intended. In postproduction, videos were temporally aligned with audio recordings such that beat gesture stroke onsets occurred exactly 200 ms prior to color word onsets, resulting in apices that co-occurred with stressed syllables of color words, consistent with the timing of gesture production relative to natural spoken discourse (Morrel-Samuels & Krauss, 1992) and with perceptual biases for gesture relative to speech (Leonard & Cummins, 2011). Videos were then cropped to the length of audio recordings. A white circular reverse mask was then added to videos so that they appeared to have a circular frame against the white background used in the experimental paradigm. This ensured that all objects in an array were equidistant from the video, which was located centrally in visual displays (similar to the implementation of video in the visual world display used in Silverman et al., 2010; see Fig. 1).

2.1.3.6. Critical sentences: For each critical sentence (2a–b, 4a–b), two videos were recorded consecutively: one in which the talker produced a beat gesture in conjunction with the color word, and one in which the talker did not produce a beat gesture. Notably, beat gesture occurred only during the color word within color-+shape-difference critical referring expressions, ensuring its presence and timing were identical during color-contrast and color-+shape-difference sentences. Thus, like contrastive accenting, the presence of beat gesture was locally felicitous in color-contrast but not color-+shape-difference referring expressions, whereas the absence of beat gesture was locally felicitous in color-+shape-difference but not color-contrast referring expressions. Note that beat gesture was factorially paired with both contrastive and non-contrastive accenting
during both recording and presentation of stimuli. Although it is possible that beat gesture produced in conjunction with contrastive accenting may have differed from beat gesture produced in conjunction with non-contrastive accenting, recording separate videos with beat gesture for critical sentences with contrastive and non-contrastive accenting allowed us to optimize gesture-accent duration synchrony and ecological validity while also permitting us to manipulate beat gesture orthogonal to pitch accenting in these trials, crucial to our experimental design.

2.1.3.7. Filler sentences: In contrast with critical sentences, a single video was recorded to accompany each filler sentence (2c–d, 4c–d). In videos accompanying shape-contrast sentences, the talker produced a beat gesture in conjunction with the shape word, whereas in videos for neither-difference sentences, the talker did not produce a beat gesture. This configuration maintained the pattern of usage of, and association between, pitch accenting and beat gesture found in natural spoken discourse, such that these cues were globally felicitous.

2.1.3.8. Norming: We conducted norming to confirm that the co-occurrence of beat gesture and contrastive accenting and their local felicity with contrast in critical and filler sentences were perceived as expected. Seventy-eight participants, who did not participate in either Experiment 1 or 2, watched a sample of stimuli from Experiments 1 and 2 and rated each of them on a 1 (completely unnatural)–7 (completely natural) basis for (a) gesture-accent co-occurrence (How well did the speaker’s gestures match the speech?) and local felicity of these cues with contrast (How well did the speech style fit the instructions given?). The co-occurrence of beat gesture and contrastive accenting was rated as more natural than either cue in isolation. Further, both cues were perceived as more natural when they were locally felicitous with contrast. Taken together, these findings confirm that beat gesture and contrastive accenting are perceived as most natural when they co-occur with one another and are locally felicitous with contrast. The full results of this norming study can be found in §S1 of the Supplementary Results section.

2.1.4. Procedure

Prior to beginning the experimental task, participants were seated 55–65 cm from the monitor on which stimuli were presented (35°55′0.32″ visual angle). Gaze was calibrated to within 0.5° of visual angle using 13 points of reference. Drift checks were performed between experimental trial blocks, and recalibration was performed if gaze was misaligned by more than 2° of visual angle.

At the beginning of the experimental task, participants were told that its purpose was to test their ability to follow instructions. Consequently, they were instructed to respond to all instructions as quickly and accurately as possible by using a mouse to click on the appropriate shape. They were also instructed that, if they accidentally clicked on the wrong shape, they would need to click on the correct shape to proceed; however, in our dataset, all responses to critical referring expressions were correct. No instructions concerning gaze were provided.
Participants first completed an eight-trial practice block to become familiar with the task, then proceeded to the experimental blocks (4 blocks of 40 trials each; see Table 1). In both practice and experimental blocks, critical and filler trials were randomly interleaved. Each trial began with a context sentence accompanied by a visual display consisting of a centrally located video and a surrounding array of objects corresponding to that sentence and the subsequent continuation sentence (see Fig. 1). Following a correct response, the video disappeared and was replaced by a gray circular placeholder for 1,000 ms while the object array remained on screen. Subsequently, the sequence repeated with the continuation sentence and its corresponding video. Following a correct response, the trial ended, and, after a blank screen was displayed for 1,000 ms, a new trial began or the block ended. After each experimental block, participants received the opportunity to take a brief break. Once they indicated that they were ready to begin the next block, a drift check was performed and, if necessary, gaze was recalibrated prior to the first trial of that block.

During experimental trials, participants were permitted to gaze freely at the monitor while gaze data were collected remotely from the right eye at a 500 Hz sampling rate using an EyeLink 1000 eye-tracker (SR Research). If participants’ gaze left the trackable range, an audible alert sounded and participants were repositioned if necessary.5

2.1.5. Pupil preprocessing

The TEPR was preprocessed using functions from the GazeR package (Geller, Winn, Mahr, & Mirman, 2020) in R v. 3.5.2. Because response accuracy for critical referring expressions was 100%, no TEPRs were excluded due to incorrect responses. Trials with greater than 30% missing pupil data were excluded (<0.01% of the data), and pupil samples three standard deviations above the mean diameter of each trial were removed (3.2% of the data). Blinks, along with samples that occurred 100 ms before and 100 ms after the blink, were coded as missing and linearly interpolated. We did not apply the algorithm based on median absolute deviation for eliminating artifacts that may arise from quick changes in pupil size, as suggested by Kret and Sjak-Shie (2019), because it would have discarded 96.7% of the deblinked pupil data. To smooth the pupil time course, a 5-point moving average was passed over the deblinked pupil data. Data from each trial were baseline corrected using linear scaling (see Reilly, Kelly, Kim, Jett, & Zuckerman, 2018) based on the median pupil diameter from the 200 ms prior to beat onset (t−200 to t−400 ms relative to critical word onset) during the initial portion of critical sentences and accompanying videos. Subsequent to baseline correction, pupil data were converted to standardized (z) scores to provide meaningful comparisons. A time window starting from gesture onset (t−200 ms relative to critical word onset) to 1,600 ms post-critical word onset was selected for the analysis, as this range seemed to capture the entire rise and initial fall of the TEPR across conditions and individuals (see Fig. 2); after this point, TEPR trajectories (rise, level, or fall) differed as a function of condition. To reduce computational cost given the large amount of pupillometric data, the pupillary time course during TEPRs was epoched into 200 ms time bins for analysis, averaging across 100 samples for each bin.
2.1.6. Growth curve analysis

The TEPR was analyzed using growth curve analysis (GCA; Mirman, 2014; for recent applications to TEPR analysis, see Geller et al., 2019; Kuchinsky et al., 2013; Winn et al., 2015). For the current analysis, a model that included interactions between first- and second-order polynomials and the independent variables of interest was used to analyze the TEPR time course because it was consistent with the quadratic trajectory of the data (see Fig. 2) and fit better than a model that included interactions between only the first-order polynomial and the independent variables of interest, \( \chi^2(8) = 72,125, p < .001; \) AIC\(_{\text{M1}}\) = 2,907,275; AIC\(_{\text{M2}}\) = 2,835,166. This quadratic model uses three parameters to characterize the curve. The intercept refers to the average pupil size over the full time window. The linear time term (ot1) refers to the slope of the pupillary time course, such that a positive slope indicates a sharper linear increase whereas a negative slope indicates a more gradual linear increase. And the quadratic term (ot2) provides the inflection point at which the curve bends; given the upward convex curve describing the TEPR function, negative inflection points indicate this point is more focal and occurs earlier (i.e., steeper curvature), whereas positive inflection points indicate this point is less focal and occurs later (i.e., shallower curvature). For example, a positive intercept value would indicate a...
larger overall pupil size, a positive ot1 value (slope) would indicate an increase in pupil size over time, and a positive ot2 value would indicate that the inflection point at which the pupil size begins to decrease is less focal and occurs later. Therefore, if the intercept is positive, ot1 increases over time (positive), and ot2’s point of inflection is less focal and occurs later (positive), this would reflect higher cognitive load.

2.2. Results

Preprocessed pupil data and analysis code for all growth curve models are publicly available via the following link: https://osf.io/fy6wp/. The growth curve model used to analyze the data from Experiment 1 was specified as follows:

```r
m.2.truncated <- lmer
(zscoredp ~ (ot1 + ot2) * contrast * accent * gesture + (1 + ot1+ot2 + contrast + accent +
gesture|subject), data = d.truncated, control = lmerControl(optimizer="bobyqa"),
REML = F).
```

This model included fixed effects of time terms (linear, quadratic) and all of the independent variables of interest, which included beat gesture (beat, no beat), pitch accent (contrastive, non-contrastive), contrast type (color, color + shape), and their interactions, as well as a random effect of participant. Although a single model was used to analyze the main effects of beat gesture and pitch accent as well as interactions between these factors and contrast type (see Table 3), we report these effects separately below to facilitate interpretation with respect to the hypotheses. Random slopes by participant for all independent variables, which were manipulated within subjects, were included in the model because they did not prevent it from converging and because they significantly improved its fit beyond that of the random intercepts-only model in likelihood-ratio tests ($p < .001$; Barr, Levy, Scheepers, & Tily, 2013). All fixed effects except for time terms were coded using mean-centered (Helmert) contrast coding. The model was fit in R with Laplace estimation using the lmer() function of the lme4 package (Bates, Mächler, Bolker, & Walker, 2015) with log-likelihood maximization using the BOBYQA optimizer to promote convergence. Null hypothesis significance testing was conducted using the lmerTest package (Kuznetsova, Brockhoff, & Christensen, 2017).

2.2.1. Effects of emphasis cues regardless of local felicity with referential contrasts

We first considered how beat gesture and contrastive accenting each affected comprehenders’ cognitive load, regardless of their fit to the local referential context, by examining the effects of these cues on the TEPR across contrast types (see Table 3; Fig. 2). For beat gesture, we observed significant negative main effects on both the linear and the quadratic terms, indicating that the TEPR increased less over time and was steeper in curvature when beat gesture was present (see Fig. 3). For pitch accent, no significant main effects were observed on any of the terms, indicating that the TEPR was similar in size, slope, and curvature regardless of the presence or absence of contrastive accenting. In addition, a significant negative two-way interaction between beat gesture and pitch accent was observed on both the linear and the quadratic terms. For the linear term, the negative
effect of beat gesture on TEPR increase was made even more negative when contrastive accenting was also present. Similarly, for the quadratic term, the steeper curvature produced beat gesture was steeper still when combined with contrastive accent.

These results indicate that—in a global context in which filler items supported a contrastive interpretation of beat gesture and contrastive pitch accenting—beat gesture initially taxes comprehenders’ cognitive resources, but this demand increases less over time, particularly when combined with contrastive accenting, than the absence of beat gesture. These results suggest that, in a felicitous global context, beat gesture affects referential
expression processing to a greater extent than contrastive accenting, and the combination of both cues increases comprehenders’ cognitive load less over time than each cue individually.

2.2.2. Effects of local felicity between emphasis cues and referential contrasts

We next considered how cognitive load was modulated by local felicity of prominence cues with contrast in critical referring expressions. Specifically, we considered the interactions between beat gesture, contrastive accenting, and contrast type on the TEPR (see Table 3; Fig. 2). We observed a significant positive main effect of contrast type on the linear and quadratic terms, indicating that the TEPR increased less over time and was steeper in curvature during color-contrast relative to color+shape-difference referring expressions (see Fig. 3).

Critical to our research questions, we also observed significant interactions between contrast type and beat gesture on the intercept, linear, and quadratic terms. Simple effect analyses by contrast type revealed that, during color-contrast referring expressions, the TEPR was larger ($t = 10.23$, $p < .001$), increased gradually over time ($t = -22.18$, $p < .001$), and was steeper in curvature ($t = -13.37$, $p < .001$) when beat gesture was present relative to when beat gesture was absent. During color+shape-difference referring expressions, these analyses revealed that the TEPR was even larger ($t = 22.75$, $p < .001$), increased less gradually over time ($t = -9.86$, $p < .001$), and was even steeper in curvature ($t = -20.38$, $p < .001$) when beat gesture was present relative to when beat gesture was absent.
Unlike beat gesture, contrastive accenting did not significantly interact with contrast type on any of the time terms.

Finally, we observed a positive significant three-way interaction between contrast type, contrastive accenting, and beat gesture on the quadratic term. Simple effect analyses by contrast type revealed that, during color-contrast referring expressions, the curvature of the TEPR was steeper when beat gesture was present relative to when it was absent \((t = -9.34, p < .001)\), whereas its curvature was similar regardless of whether contrastive accenting was present or absent \((t < 1)\), and the interaction between beat gesture and accent was marginal \((t = -1.65, p = .099)\). During color + shape-difference referring expressions, the same analysis revealed that the curvature of the TEPR was even steeper when beat gesture was present relative to when it was absent \((t = -16.60, p < .001)\), whereas its curvature was similar regardless of whether contrastive accenting was present or absent \((t < 1)\), and the interaction between these factors was significant \((t = 4.60, p < .001)\). Taken together, these results indicate that although the combination of beat gesture and contrastive accenting generally elicited a steeper TEPR peak, this was not so during color-contrast referring expressions, when the combination of these cues elicited a shallower TEPR peak.

Taken together, these results provide evidence that the effects of beat gesture and contrastive accenting on comprehenders’ cognitive load are modulated by felicity of these cues with the local referential context. Specifically, when both beat gesture and contrastive accenting are locally felicitous with contrast, comprehenders’ cognitive resource consumption peaks relatively late.

2.3. Discussion

We examined how two cues to contrast—beat gesture and contrastive accenting—affect comprehenders’ cognitive load during processing of referring expressions in which those cues were either locally felicitous (color-contrast) or infelicitous (color- + shape-difference). In Experiment 1, we tested these effects in a global context in which filler items contained beat gesture and contrastive accenting aligned with the shape word, such that these cues were always felicitous given the contrasts in filler referring expressions. The results revealed that, in critical referring expressions, beat gesture initially taxed comprehenders’ cognitive resources, particularly when it co-occurred with contrastive accenting and when it was locally infelicitous—that is, when beat gesture was used to emphasize the color adjective even though there was not a specific contrast in color.

Although beat gesture may initially tax comprehenders’ cognitive resources, as suggested by the steeper TEPR curvature observed during critical referring expressions with beat gesture, it may ultimately demand fewer cognitive resources than the absence of beat gesture, as evidenced by a smaller increase in TEPR slope over time. Notably, the TEPR elicited by the presence of beat gesture was smaller and increased more gradually over time when beat gesture was locally felicitous with critical referring expressions than when beat gesture was locally infelicitous with critical referring expressions. These results complement mixed findings from previous work showing that viewing relevant
representational gesture sometimes improves learning (Alibali & Nathan, 2007; Lazaron, 2004; McCafferty, 2002; Valenzeno et al., 2003) and sometimes hinders learning (Byrd et al., 2014; Yeo et al., 2017). In light of our present results, one interpretation of these previous mixed findings is that beat gesture’s demand on comprehenders’ cognitive resources may ultimately either improve or hinder discourse processing depending on whether comprehenders process language more efficiently or disengage (see Section 4.3. for further discussion of this idea).

It is worth noting that the presence of beat gesture inherently contains more visual information than its absence, so the differences in cognitive demands elicited by it may, to some extent, be due to visual processing. Nevertheless, the fact that beat gesture interacted with contrastive accenting and with the local felicity of accompanying referring expressions indicates that the effect of beat gesture on comprehenders’ cognitive load is also driven by its linguistic function. Notably, our current work provides the first evidence that beat gesture’s local felicity with contrast affects its impact on comprehenders’ cognitive load, as evidenced by the relatively smaller size, more gradual increase over time, and shallower curvature of the TEPR observed when beat gesture was present relative to when beat gesture was absent during locally felicitous than locally infelicitous critical referring expressions.

Unlike beat gesture, contrastive accenting did not affect comprehenders’ cognitive load on its own, but when it occurred in conjunction with beat gesture, it led to a smaller increase in cognitive load over time. Moreover, shallower TEPR curvature was observed when contrastive accenting and beat gesture were felicitous than when they were infelicitous with contrast in critical referring expressions. No interaction between contrastive accenting, beat gesture, and contrast type was observed for the overall size of the TEPR, however, indicating that comprehenders’ cognitive load did not differ based on these factors. These results differ from the finding of Zellin et al. (2011), obtained in an experiment in which beat gesture was never present, that felicitous contrastive accenting given linguistic contrasts reduced comprehenders’ cognitive load. This difference in results is likely due to the presence of beat gesture within the current study—particularly within a linguistic context in which contrastive accenting and beat gesture frequently co-occurred and were globally felicitous with contrast in filler referring expressions. This interpretation is consistent with our previous work, which shows whether contrastive accenting enhances subsequent memory for a spoken discourse depends in part on whether beat gesture is also present within the linguistic context (Morett & Fraundorf, 2019). It is noteworthy, however, that the initial tax on comprehenders’ cognitive load from the combination of contrastive accenting and beat gesture is inconsistent with the predictions of dual-coding theory. This inconsistency may be due to the infelicity of these cues with contrast in some critical referring expressions, which may have been particularly salient given that these cues were always felicitous with contrast in filler referring expressions. In Experiment 2, we determine whether this inconsistency persisted in a linguistic context in which contrastive accenting and beat gesture were always infelicitous with contrast in filler referring expressions.
In sum, Experiment 1 suggests that a beat gesture’s local felicity with contrast in critical referring expressions—particularly when it occurs in conjunction with contrastive accenting— influences its impact on comprehenders’ cognitive load. However, one caveat is that, even if participants had no prior expectation that beat gesture and contrastive accenting convey contrast, they could have learned this solely on the basis of the filler trials. Thus, in Experiment 2, we used a different set of fillers in which we manipulated beat gesture and pitch accent orthogonally in relation to the color word rather than jointly in relation to the shape word, such that neither beat gesture nor pitch accent was informative as to target referent identity within the global linguistic context; thus, participants could not learn how to interpret these cues merely from the experimental stimuli. Given that, in natural discourse, beat gesture and contrastive accenting would typically occur in conjunction with the shape word rather than the color word in these referring expressions, the filler referring expressions of Experiment 1 provided a more ecologically valid—albeit predictable—context, whereas those of Experiment 2 provided an unpredictable—albeit less ecologically valid—context, complementing one another.

3. Experiment 2: Infelicitous context

In Experiment 2, we again examined the effects of beat gesture and contrastive accenting on cognitive load during spoken discourse comprehension, but now within a context in which these cues were globally infelicitous, controlling for possible within-experiment learning that may have been present in Experiment 1.

3.1. Methods

3.1.1. Participants

Forty adult monolingual native English speakers (age range: 18–26 years; 36 women, 4 men) separate from those who participated in Experiment 1 participated in Experiment 2 for partial course credit. All participants were recruited from a large public research university in the Southeastern United States via relevant courses. All participants reported normal hearing, normal or corrected-to-normal vision, and no color blindness.

3.1.2. Design

As in Experiment 1, the orthogonal manipulation of contrast type, pitch accent, and beat gesture in critical referring expressions resulted in a 2 (color-contrast vs. color- + shape-difference) × 2 (contrastive vs. non-contrastive accenting) × 2 (beat vs. no beat) within-participants design. Once again, eight lists counterbalanced by contrast type, pitch accent, and beat gesture (Lists 1.1–2.4) were used for both practice trials and experimental and filler trials. As in Experiment 1, because beat gesture and pitch accenting were manipulated on color words, their presence was locally felicitous in the color-contrast condition and locally infelicitous in the color- + shape-difference condition,
whereas the opposite was true for their absence. Fig. 4 displays the sentences and accompanying visual arrays used in the first critical and filler trials of Lists 1.1–1.4. (See Table S6 in the Supplementary Material for additional information concerning trials presented in each list.)

3.1.3. Materials
3.1.3.1. Objects: The same 64 objects used in Experiment 1 were used in Experiment 2. Objects presented during critical referring expressions were assigned in the same way as in Experiment 1. As in Experiment 1, during all trials, objects appeared in one of four locations equidistant from centrally presented, circularly framed videos (see Fig. 3). Again, for each item, the positioning of objects was counterbalanced across participants such that context, target, competitor, and distractor objects were equally likely to appear in each position (see Tables S4 and S7 in the Supplementary Material).

3.1.3.2. Audio recordings: The same spliced audio recordings of critical sentences created for Experiment 1 were used as critical sentences in Experiment 2. To prevent participants from learning when and how beat gesture and contrastive accenting would occur, new filler sentences were created in which these cues were manipulated orthogonally on the color word, as they were in critical sentences. To create the new filler sentences, we spliced color words with contrastive and non-contrastive accenting from critical sentences into carrier sentences that we created from filler sentences in Experiment 1 with the same combinations of color and shape words as new filler sentences. Although the number and ordering of filler sentences with specific color and shape word combinations differed between Experiments 1 and 2 due to this splicing, the number of sentences with each contrast type and accent in the practice and experimental blocks for each set of lists was the same in both experiments.

As in Experiment 1, half of filler referring expressions differed in shape word but not in color word relative to the preceding context (shape-contrast), and half differed in
neither shape word nor color word (neither-difference). However, in Experiment 2 fillers, pitch accenting differed in color words. Because the targets of fillers were always the same color as the preceding context object, the presence of contrastive accenting on color words was always infelicitous, such that pitch accenting was globally infelicitous (see Table S5 in the Supplementary Material). In order to distribute all combinations of beat gesture and pitch accent evenly across sentences with all contrast types, context and critical sentences were paired differently and trials were ordered differently in critical and filler trials than they were in Experiment 1 (see Appendix C, publicly available with other appendices at https://osf.io/fy6wp/, for complete list of sentences used in Experiment 2).

3.1.3.3. Video recordings: The same videos used with critical sentences in Experiment 1 were used with critical sentences in Experiment 2. To construct additional videos to accompany the new filler sentences constructed for Experiment 2, we temporally realigned video recordings used with Experiment 1 fillers to match the different audio recordings used in Experiment 2 fillers. (See Table S8 in the Supplementary Material for the number of videos constructed for each set of lists.) Videos were temporally realigned with audio recordings of filler sentences, such that beat gesture stroke onsets (temporally realigned in Experiment 1 to occur 200 ms prior to shape word onsets in fillers) occurred 200 ms prior to color word onsets, resulting in apices co-occurring with stressed syllables of color words. Videos without beat gesture were similarly temporally realigned with audio recordings of filler sentences, such that shape words in audio tracks of videos (prior to removal) co-occurred with color words in audio recordings. This configuration maintained the temporal association between pitch accenting and beat gesture, but—by design—eliminated the temporal association between these cues and contrasting information (shape words) found in natural spoken discourse, such that these cues were globally infelicitous. Temporal realignment resulted in videos shorter in length than the corresponding audio recordings for Experiment 2 fillers. To fill this gap in timing, the last frame of each video was extended over its duration, resulting in a freeze frame at the end of each video for fillers. 8

3.1.3.4. Procedure: The same procedure used in Experiment 1 was used in Experiment 2.

3.1.3.5. Pupil preprocessing: The TEPR was preprocessed using the same procedure used in Experiment 1.

3.1.3.6. Growth curve analysis: As in Experiment 1, changes in pupil diameter were analyzed using GCA (Mirman, 2014). Again, a model that included interactions between first- and second-order polynomials and the independent variables of interest was used to analyze TEPR time course because it was consistent with the quadratic trajectory of the data (see Fig. 5 below) and fit better than a model that included interactions between only the first-order polynomial and the independent variables of interest, \( \chi^2(8) = 12,998, p < .001; \text{AIC}_M = 2,421,695; \text{AIC}_{M2} = 2,408,713. \)
3.2. Results

Preprocessed pupil data and analysis code for all growth curve models are publicly available via the following link: https://osf.io/fy6wp/. The growth curve model used to analyze the data from Experiment 2 was specified as follows:

\[
m.2.\text{truncated} \leftarrow \text{lmer}(\text{zscoredp} \sim (\text{ot1} + \text{ot2}) \ast \text{contrast} \ast \text{accent} \ast \text{gesture} + (1 + \text{ot1} + \text{ot2} + \text{contrast} + \text{accent} + \text{gesture}|\text{subject}), \text{data} = \text{d.truncated}, \text{control} = \text{lmerControl(optimizer = "bobyqa"), REML = F}).
\]

As in Experiment 1, this model included fixed effects of time terms (linear, quadratic) and the independent variables of interest, which included beat gesture (beat, no beat), pitch accent (contrastive, non-contrastive), contrast type (color, color + shape), and their interactions, as well as a random effect of participant. Although a single model was used to analyze the main effects of beat gesture and pitch accent as well as the main effect of contrast type and its interactions with beat gesture and pitch accent (see Table 4), we again report these effects separately below to facilitate interpretation. Again, random slopes by participant for all of these independent variables, which were manipulated...
within subjects, were included in the model because they did not prevent it from converging and because they significantly improved its fit beyond that of the random intercepts-only model in likelihood-ratio tests ($p < .001$; Barr et al., 2013).

Because paradigms and participants differed between Experiments 1 and 2, data from each experiment were analyzed and are reported separately here. Because the designs of the experiments are largely similar except for beat gesture and pitch accent felicity in filler sentences, however, a combined analysis that includes experiment as a fixed effect to directly test the effect of filler felicity is available in Section S2 of the Supplementary Results section.

<table>
<thead>
<tr>
<th>Fixed Effect</th>
<th>Coefficient</th>
<th>SE</th>
<th>t</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>0.01</td>
<td>0.03</td>
<td>0.04</td>
<td>.97</td>
</tr>
<tr>
<td>ot1</td>
<td>0.27</td>
<td>0.08</td>
<td>3.47</td>
<td>&lt;.001***</td>
</tr>
<tr>
<td>ot2</td>
<td>-0.40</td>
<td>0.04</td>
<td>-9.70</td>
<td>&lt;.001***</td>
</tr>
<tr>
<td>Beat gesture (vs. no gesture)</td>
<td>0.03</td>
<td>0.03</td>
<td>0.73</td>
<td>.46</td>
</tr>
<tr>
<td>ot1 $\times$ beat gesture</td>
<td>0.11</td>
<td>0.01</td>
<td>11.4</td>
<td>&lt;.001***</td>
</tr>
<tr>
<td>ot2 $\times$ beat gesture</td>
<td>0.06</td>
<td>0.01</td>
<td>6.7</td>
<td>&lt;.001***</td>
</tr>
<tr>
<td>Contrastive accenting (vs. non-contrastive)</td>
<td>0.02</td>
<td>0.04</td>
<td>0.45</td>
<td>.65</td>
</tr>
<tr>
<td>ot1 $\times$ contrastive accenting</td>
<td>0.1</td>
<td>0.01</td>
<td>10.04</td>
<td>&lt;.001***</td>
</tr>
<tr>
<td>ot2 $\times$ contrastive accenting</td>
<td>0.03</td>
<td>0.01</td>
<td>3.31</td>
<td>&lt;.001***</td>
</tr>
<tr>
<td>Beat gesture $\times$ contrastive accenting</td>
<td>-0.04</td>
<td>0.01</td>
<td>-6.2</td>
<td>&lt;.001***</td>
</tr>
<tr>
<td>ot1 $\times$ beat gesture $\times$ contrastive accenting</td>
<td>0.15</td>
<td>0.02</td>
<td>7.61</td>
<td>&lt;.001***</td>
</tr>
<tr>
<td>ot2 $\times$ beat gesture $\times$ contrastive accenting</td>
<td>-0.1</td>
<td>0.02</td>
<td>-5.32</td>
<td>&lt;.001***</td>
</tr>
<tr>
<td>Contrast type (color vs. color + shape)</td>
<td>0.02</td>
<td>0.04</td>
<td>0.5</td>
<td>.62</td>
</tr>
<tr>
<td>ot1 $\times$ contrast type</td>
<td>0.01</td>
<td>0.01</td>
<td>1.69</td>
<td>.09†</td>
</tr>
<tr>
<td>ot2 $\times$ contrast type</td>
<td>0.08</td>
<td>0.01</td>
<td>11.71</td>
<td>&lt;.001***</td>
</tr>
<tr>
<td>Beat gesture $\times$ contrast type</td>
<td>-0.01</td>
<td>0.01</td>
<td>-1.62</td>
<td>.11</td>
</tr>
<tr>
<td>ot1 $\times$ beat gesture $\times$ contrast type</td>
<td>-0.2</td>
<td>0.01</td>
<td>-14.47</td>
<td>&lt;.001***</td>
</tr>
<tr>
<td>ot2 $\times$ beat gesture $\times$ contrast type</td>
<td>-0.01</td>
<td>0.01</td>
<td>-0.24</td>
<td>.81</td>
</tr>
<tr>
<td>Contrastive accenting $\times$ contrast type</td>
<td>-0.01</td>
<td>0.01</td>
<td>-2.7</td>
<td>.007**</td>
</tr>
<tr>
<td>ot1 $\times$ contrastive accenting $\times$ contrast type</td>
<td>-0.07</td>
<td>0.01</td>
<td>-5.21</td>
<td>&lt;.001***</td>
</tr>
<tr>
<td>ot2 $\times$ contrastive accenting $\times$ contrast type</td>
<td>0.06</td>
<td>0.01</td>
<td>4.71</td>
<td>&lt;.001***</td>
</tr>
<tr>
<td>Beat gesture $\times$ contrastive accenting $\times$ contrast type</td>
<td>-0.24</td>
<td>0.01</td>
<td>-27.09</td>
<td>&lt;.001***</td>
</tr>
<tr>
<td>ot1 $\times$ beat gesture $\times$ contrastive accenting $\times$ contrast type</td>
<td>-0.22</td>
<td>0.03</td>
<td>-7.94</td>
<td>&lt;.001***</td>
</tr>
<tr>
<td>ot2 $\times$ beat gesture $\times$ contrastive accenting $\times$ contrast type</td>
<td>0.32</td>
<td>0.03</td>
<td>11.98</td>
<td>&lt;.001***</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Random Effect</th>
<th>$s^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Participant</td>
<td>0.18</td>
</tr>
<tr>
<td>Participant $\times$ ot1</td>
<td>0.42</td>
</tr>
<tr>
<td>Participant $\times$ ot2</td>
<td>0.22</td>
</tr>
<tr>
<td>Participant $\times$ beat gesture</td>
<td>0.19</td>
</tr>
<tr>
<td>Participant $\times$ contrastive accenting</td>
<td>0.22</td>
</tr>
<tr>
<td>Participant $\times$ contrast type</td>
<td>0.23</td>
</tr>
</tbody>
</table>

$p < .1$; $**p < .01$; $***p < .001$. 

Table 4

Fixed effect (top) and variance estimates (bottom) for multilevel model of comprehender pupil size from 200 ms before to 1,600 ms after color word onset in Experiment 2 (observations = 822,986)
3.2.1. Effects of emphasis cues regardless of local felicity with referential contrasts

As in Experiment 1, we first considered how beat gesture and contrastive accenting each affected comprehenders’ cognitive load, regardless of their fit to the local referential context, by examining the effects of these cues on the TEPR across contrast types (see Table 4; Fig. 5). In contrast to the effects of beat gesture and contrastive accenting observed in Experiment 1, we observed significant positive main effects of both of these cues on the linear and quadratic terms, indicating that the TEPR increased more over time and was shallower when beat gesture and contrastive accenting were present relative to when they were absent (see Fig. 6). These main effects were qualified by significant two-way interactions between beat gesture and contrastive accenting on the intercept, linear, and quadratic terms. Simple effect analyses by beat gesture revealed that, when beat gesture was present, the TEPR was smaller \((t = -20.68, p < .001)\), increased more over time \((t = 5.64, p < .001)\), and was shallower in curvature \((t = 7.55, p < .001)\) when contrastive accenting was present than when it was absent. When beat gesture was absent, these analyses revealed that the TEPR was larger \((t = 34.12, p < .001)\), increased more over time \((t = 7.11, p < .001)\), and was shallower in curvature \((t = 3.83, p < .001)\) when contrastive accenting was present than when it was absent. Together, these results indicate that, in a context in which beat gesture and contrastive accenting are globally infelicitous in filler items, the presence of both of these cues in critical referring expressions decreased overall cognitive load, whereas contrastive accenting unaccompanied by beat gesture and beat gesture unaccompanied by contrastive accenting in critical referring expressions increased overall cognitive load.

Fig. 6. Decomposition by time terms of comprehender pupil size mean (z-score) by beat gesture and pitch accent (C: contrastive; NC: non-contrastive) by contrast type (color; color + shape) from 200 ms before to 1,600 ms after color word onset in Experiment 2.
Taken together, these results indicate that—in a global linguistic context in which filler items did not support a contrastive interpretation of beat gesture and contrastive pitch accenting—beat gesture and contrastive accenting each demanded cognitive resources during spoken language comprehension. However, beat gesture mitigated the effect of contrastive accenting on comprehenders’ overall cognitive load. Taken together, these results suggest that, in a globally infelicitous linguistic context, the combination of beat gesture and contrastive accenting may cause comprehenders to disengage, hindering referential expression processing (see Section 4.3 for further discussion of this idea).

3.2.2. Effects of local felicity of emphasis cues with referential contrast

We next considered how the match between cues to prominence and the content of referring expressions affects comprehenders’ cognitive load by examining interactions between beat gesture, contrastive accenting, and contrast type (see Table 4; Fig. 5). For contrast type, we observed a marginal positive main effect on the linear term and a significant positive main effect on the quadratic term, indicating that the TEPR increased slightly less over time and was steeper in curvature during color-contrast relative to color- + shape-difference referring expressions (see Fig. 6).

Critical to our research question, we observed a negative two-way interaction between contrast type and beat gesture on the linear term. Simple effect analyses by contrast type revealed that, during color-contrast referring expressions, the TEPR increased more over time when beat gesture was present than when it was absent ($t = 7.36, p < .001$), whereas during color- + shape-difference referring expressions, the TEPR increased less over time when beat gesture was present relative to when it was absent ($t = -8.34, p < .001$). This crossover interaction can also be seen in simple effect analyses by beat gesture, which similarly revealed that, when beat gesture was present, the TEPR increased more over time during color-contrast relative to color- + shape-difference referring expressions ($t = 8.65, p < .001$), whereas when beat gesture was absent, the TEPR increased less during color-contrast relative to color- + shape-difference referring expressions ($t = 14.84, p < .001$).

Additionally, we also observed two-way interactions between contrast type and contrastive accenting on the intercept, linear, and quadratic terms. Simple effect analyses by contrast type revealed that, during color-contrast referring expressions, the TEPR was larger ($t = 7.20, p < .001$), increased more over time ($t = 11.01, p < .001$), and was shallower in curvature ($t = 5.00, p < .001$) when contrastive accenting was present relative to when it was absent. By contrast, during color + shape-difference referring expressions, these analyses revealed that the TEPR was similar in size ($t < 1$), similar in slope ($t < 1$), and was shallower in curvature ($t = 8.09, p < .001$) when contrastive accenting was present relative to when it was absent.

Finally, we observed significant three-way interactions between contrast type, contrastive accenting, and beat gesture on the intercept, linear, and quadratic terms. Simple effect analyses by contrast type revealed that, during color-contrast referring expressions, the TEPR was larger ($t_{\text{Gesture}} = 10.66, p_{\text{Gesture}} < .001$; $t_{\text{Accent}} = 6.56, p_{\text{Accent}} < .001$), increased more sharply over time ($t_{\text{Gesture}} = 11.48, p_{\text{Gesture}} < .001$; $t_{\text{Accent}} = 10.73, p_{\text{Accent}} < .001$),
$p_{\text{Accent}} < .001$), and was shallower in curvature ($t_{\text{Gesture}} = 6.95, p_{\text{Gesture}} < .001$; $t_{\text{Accent}} = 4.37, p_{\text{Accent}} < .001$) when beat gesture and contrastive accenting were each present relative to when they were each absent. Moreover, during color-contrast referring expressions, the TEPR was smaller ($t = -9.54, p < .001$), increased more sharply over time ($t = 9.65, p < .001$), and was steeper in curvature ($t = -3.38, p < .001$) when beat gesture and contrastive accenting occurred in combination relative to when they each occurred alone. By comparison, during color + shape-difference referring expressions, the TEPR was larger ($t = 3.04, p = .002$), increased more gradually over time ($t = -9.64, p < .001$), and was shallower in curvature ($t = 5.50, p < .001$) when beat gesture was present relative to when it was absent, whereas it was similar in size ($t < 1$), showed a nonsignificant trend toward increasing more sharply over time ($t = 1.70, p = .09$), and was shallower in curvature ($t = 9.33, p < .001$) when contrastive accenting was present relative to when it was absent. Moreover, the TEPR was smaller ($t = -39.70, p < .001$), showed a nonsignificant trend toward increasing more gradually ($t = -1.79, p = .07$), and was shallower in curvature ($t = 12.23, p < .001$) when beat gesture and contrastive accenting occurred in combination relative to when they occurred alone. Together, these results indicate that the super-additive effect of the presence of both cues on the increase in comprehenders’ cognitive load over time was reversed during color-contrast referring expressions. Moreover, they indicate that the sub-additive effects of beat gesture and pitch accenting on TEPR curvature became super-additive during color-contrast referring expressions, yielding an especially steep curve.

To summarize, Experiment 2 presented a different global linguistic context—one in which filler items did not support a contrastive interpretation of beat gesture and contrastive pitch accenting. Within this context, cues that were locally infelicitous with critical referring expressions (color- + shape difference trials) each increased cognitive load more on their own but increased it less in combination with the other cue. However, no such pattern emerged when these cues were locally felicitous with critical referring expressions (color-contrast). Taken together, these findings provide evidence that global as well as local felicity of beat gesture and contrastive accenting with contrast affects comprehenders’ cognitive resource consumption.

3.3. Discussion

In Experiment 2, we again examined how beat gesture and contrastive accenting affect comprehenders’ cognitive load during processing of locally felicitous (color-contrast) or locally infelicitous (color- + shape-difference) critical referring expressions. The critical difference is that these variables were now tested in a global linguistic context in which the presence of these cues was infelicitous with contrast in filler referring expressions. We found that beat gesture and contrastive accenting each increased cognitive load, particularly when they were locally infelicitous with contrast in critical referring expressions. Moreover, we found that the combination of these cues decreased overall cognitive load, particularly when they were locally infelicitous with contrast in critical referring expressions.
Unlike in Experiment 1, in which beat gesture initially taxed but increased comprehenders’ cognitive load less over time, in Experiment 2, we found that beat gesture and contrastive accenting each failed to initially tax comprehenders’ cognitive load but increased it more over time. This difference in cognitive demands was likely due to the global infelicity of these cues with contrast in filler referring expressions of Experiment 2, which interfered with the cues’ utility in predicting contrast in critical referring expressions. This explanation is consistent with the finding that these effects were larger when beat gesture and contrastive accenting were locally infelicitous than when they were locally felicitous with contrast in critical referring expressions. These findings complement evidence from previous work that comprehenders leverage contrastive accenting to anticipate upcoming contrasts, especially when these cues occur in felicitous linguistic contexts (Ito & Speer, 2008; Kurumada, Brown, Bibly, Pontillo, & Tanenhaus, 2014; Watson et al., 2008), thereby reducing the cognitive load of spoken discourse comprehension. Specifically, the increase in cognitive load over time elicited by beat gesture and contrastive accenting in Experiment 2 indicates that global *infelicity* of contrastive accenting and beat gesture with contrast increases comprehenders’ cognitive load during reference resolution.

The finding that the combination of beat gesture and contrastive accenting mitigates overall demand on comprehenders’ cognitive resources provides evidence that the effects of these cues on comprehenders’ cognitive load are interactive rather than strictly independent, particularly when these cues are locally infelicitous with contrast in co-occurring spoken discourse. This finding complements previous research in which these cues were always felicitous with contrastive information in spoken discourse, the results of which revealed that beat gesture influences the effect of contrastive accenting on memory for this information (Kushch & Prieto, 2016; Llanes-Coromina et al., 2018; Morett & Fraundorf, 2019). Although it is not entirely clear from the pupillometric data why beat gesture reduces comprehenders’ cognitive load during interpretation of infelicitous referring expressions, it may be because a consistent lack of felicity with contrast (in both filler and locally infelicitous critical referring expressions) encourages comprehenders to tune out and expend fewer cognitive resources during processing of locally infelicitous referring expressions. This explanation remains speculative, however, and should be verified with additional research.

### 4. General discussion

In the current research, we investigated how contrastive accenting and beat gesture affect comprehenders’ cognitive load in relation to their local and global felicity with contrast during online spoken referring expression processing. In particular, we investigated whether these cues decrease comprehenders’ cognitive load when they co-occur, as postulated by dual-coding theory (Paivio, 1990); increase cognitive load when they co-occur, as postulated by additive theories of cue processing (e.g., Just & Carpenter, 1992); or affect cognitive load in accordance with their co-occurrence with each other and linguistic contrast (Crosse et al., 2015; Olasagasti et al., 2015). The results indicate that beat
gesture—and, in Experiment 2, contrastive accenting—increases cognitive load, particularly when these cues are locally infelicitous with contrast. Moreover, they indicate that, within a context in which beat gesture and contrastive accenting are globally felicitous with referential contrast and each other (Experiment 1), co-occurrence of these cues increases their impact on comprehenders’ cognitive load when they are locally felicitous—but not locally infelicitous—with referential contrast. In contrast, within a context in which beat gesture and contrastive accenting are globally infelicitous with referential contrast and each other (Experiment 2), co-occurrence of these cues decreases their impact on comprehenders’ cognitive load when they are locally infelicitous—but not locally felicitous—with referential contrast. Although these results do not align clearly with either dual-coding theory or additive theories of cue integration, they nevertheless provide evidence that comprehenders are sensitive to both local and global conflicts between beat gesture and contrastive accenting and referential contrast, as evidenced by differences in cognitive resource consumption, consistent with the predictions of growth point theory (McNeill, 1992, 2005). Moreover, these results are consistent with predictive coding accounts of audiovisual cue integration (Crosse et al., 2015; Olasagasti et al., 2015), which postulate that auditory and visual cues are processed and integrated differently, in turn affecting cognitive resource consumption, based on contextual congruency. We discuss possible explanations and theoretical ramifications of these findings below.

4.1. Local and global felicity of referential cues

The key finding of the current research is that the effects of beat gesture and contrastive accenting—both alone and in combination—on comprehenders’ cognitive load were influenced by felicity of these cues with contrast in discourse. Specifically, when beat gesture and contrastive accenting were globally felicitous with contrast and one another (Experiment 1, in which beat gesture and contrastive accenting were always used felicitously in filler items), we observed that (a) beat gesture increased comprehenders’ cognitive load, particularly in conjunction with contrastive accenting; and (b) processing of beat gesture both alone and in combination with contrastive accenting demanded more cognitive resources when it was locally felicitous than when it was locally infelicitous with contrast. On the other hand, when beat gesture and contrastive accenting were globally infelicitous with contrast and one another ( Experiment 2, in which the presence of beat gesture and contrastive accenting was always infelicitous in filler items), we found that (a) beat gesture and contrastive accenting independently increased but conjointly decreased comprehenders’ cognitive load; and (b) these effects were larger when these cues were locally infelicitous than when they were locally felicitous with contrast. Taken together, these findings suggest that when beat gesture and contrastive accenting are used inappropriately, comprehenders disengaged when these cues occurred conjointly in critical items, particularly when these cues were locally infelicitous with contrast (see Section 4.3. for further discussion).

Taken together, these findings provide the first evidence that beat gesture’s local and global felicity with contrast influences comprehenders’ cognitive resource consumption.
As such, it complements a growing body of evidence that comprehenders base their interpretation of cues to contrast on their situation-specific validity (Grodner & Sedivy, 2011; Morett & Fraundorf, 2019; Ryskin et al., 2019). Moreover, it provides an explanation for why we obtained different results than previous research showing that contrastive accenting reduces comprehenders’ cognitive load (Zellin et al., 2011): In that work, contrastive accenting was never locally or globally infelicitous with contrast. Indeed, it is possible that the absence of an experiment in which beat gesture and contrastive accenting were always felicitous given the discursive context may have affected discourse processing in the current research, so it will be important to determine whether the results replicate in a more ecologically valid linguistic context in which these cues are always produced felicitously.

4.2. Comparing beat gesture and contrastive accenting

In general, the results of the current research indicate that, during resolution of contrastive referring expressions, beat gesture and contrastive accenting exert differing demands on comprehenders’ cognitive resources. Specifically, when beat gesture and contrastive accenting were always globally felicitous in filler items (Experiment 1), contrastive accenting did not affect comprehenders’ cognitive load during critical referring expression processing, whereas beat gesture increased it, particularly when it was locally felicitous with contrast. Moreover, when beat gesture and contrastive accenting were globally infelicitous in filler items (Experiment 2), beat gesture and contrastive accenting each increased comprehenders’ cognitive load overall during critical referential expression processing, but their effects differed in direction based on local felicity with contrast. These results complement the findings of other work indicating that beat gesture enhances memory for contrasts in discourse when it co-occurs with contrastive accenting, but not when it occurs in the absence of contrastive accenting (Kushch & Prieto, 2016; Llanes-Coromina et al., 2018; Morett & Fraundorf, 2019). Together, the findings of the current research and this previous work suggest that beat gesture (a) is interpreted as a cue to contrast mainly when it is felicitous with contrast in discourse or occurs conjointly with contrastive accenting; and (b) bolsters contrastive interpretation of contrastive accenting.

4.3. Theoretical ramifications

The results of the current work were inconsistent with the hypothesis, based on dual-coding theory (Paivio, 1990), that beat gesture and contrastive accenting would decrease comprehenders’ cognitive load when they occurred conjointly. We observed that beat gesture and contrastive accenting conjointly increased cognitive load comparatively less only in a linguistic context in which these cues were globally infelicitous with contrast in filler items (Experiment 2) when they were also locally infelicitous with contrast in critical items. Because cognitive resource consumption decreased when beat gesture and contrastive accenting were locally and globally infelicitous with contrast, it likely reflects disengagement rather than facilitated language processing, suggesting that comprehenders tend to “tune out” when these cues are uninformative. Conversely, we observed that beat
gesture and contrastive accenting conjointly increased cognitive load in a linguistic context in which these cues were globally felicitous with contrast in filler items (Experiment 1), particularly when they were also locally felicitous with contrast in critical items. This finding suggests that beat gesture and contrastive accenting affect cognitive resource consumption in contexts in which these cues are informative, particularly when they are locally felicitous given the context.

The results of the current work also failed to unequivocally support the alternate hypothesis, based on additive theories of cue processing, that each additional cue or constraint on linguistic processing consumes additional cognitive resources (Just & Carpenter, 1992). In particular, the findings that (a) contrastive accenting alone increased cognitive load when it and beat gesture were globally infelicitous with contrast (Experiment 2), but not when these cues were globally felicitous with contrast (Experiment 1), and (b) beat gesture and contrastive accenting conjointly decreased cognitive load when these cues were globally felicitous with contrast (Experiment 1), are inconsistent with additive theories of cue integration. Thus, the results must be contextualized in terms of other theoretical frameworks.

One type of theory that provides a plausible account of the results is predictive coding frameworks of audiovisual cue integration (Crosse et al., 2015; Olasagasti et al., 2015). These frameworks postulate that temporal and contextual congruency of audiovisual cues with speech affects integration such that cue congruency differentially impacts cognitive processing during language comprehension. Unlike dual-coding theory and additive theories, predictive coding frameworks account for differences in how cues from the visual and verbal modalities are processed based on their frequency of co-occurrence with one another and other information, which in turn affects how they are integrated with speech. Given that predictive coding frameworks have exclusively been applied to speech perception to date, they may not yet be sufficient to explain our results concerning reference resolution, particularly with respect to local and global infelicity of cues with contrast. Thus, they may need to be supplemented by growth point theory (McNeill, 2005), which postulates that gesture and speech arise from unitary semantic representations and that semantic incongruency of gesture with co-occurring speech disrupts language comprehension. Together, predictive coding frameworks and growth point theory provide a powerful theoretical account for the findings of the current research, providing insight into how integrating beat gesture and contrastive accenting with contrast affects comprehenders’ cognitive load during reference resolution within spoken discourse comprehension.

4.4. Limitations and future directions

Although the present research provides important insight into how beat gesture and contrastive accenting affect comprehenders’ cognitive load during processing of contrasting and non-contrasting spoken referring expressions, it has some limitations. First, non-controlled factors (e.g., metrical patterns) in color words (e.g., red vs. orange) may have influenced pitch accent production and perception in shape words, which may explain why acoustic markers of differences in pitch accent were less reliable in shape than in
Second, only the talker’s hands and upper torso were visible, possibly exaggerating the effect of beat gesture on comprehenders’ cognitive load relative to more naturalistic conversational contexts in which the face and other body parts are visible. To determine the extent to which the results for beat gesture generalize to such contexts, this research ideally should be replicated with videos featuring the talker’s entire body, including the face. Third, because recent work has demonstrated that beat gesture is defined based on temporal synchrony with speech prosody rather than form (Prieto, Cravotta, Kushch, Rohrer, & Vilà-Giménez, 2018; Shattuck-Hufnagel, Ren, Mathew, Yuen, & Demuth, 2016), we viewed it as infeasible to include a control condition featuring hand movements identical in magnitude and timing to those in the beat gesture condition that would not be interpreted as beat gestures. Thus, it may be impossible to differentiate between the effect of beat gesture in particular and additional visual information in general. Finally, as mentioned above, this research included cases in which beat gesture and contrastive accenting were infelicitous with contrast in both critical items (Experiments 1 and 2) and filler items (Experiment 2), whereas such cases are rare in naturalistic discourse. Future research should investigate how beat gesture and contrastive accenting are integrated and how they affect comprehenders’ cognitive load in linguistic contexts in which they are never infelicitous with contrast, providing further insight into how such infelicity affects these cues’ impact on cognitive load during spoken discourse comprehension.

4.5. Conclusion

Overall, the results of the current research provide evidence that beat gesture increases comprehenders’ cognitive load during online spoken discourse processing, particularly when it is locally infelicitous with contrast in critical discourse or globally infelicitous with contrast in filler discourse. Moreover, they suggest that beat gesture conveys contrast via felicity with contrastive information in discourse or co-occurrence with other cues that convey contrast, such as contrastive accenting. Notably, these results provide the first evidence that the TEPR can be used to examine how beat gesture—and contrastive accenting—affects comprehenders’ cognitive load implicitly in real time, providing insight into how multimodal cues affect the moment-to-moment effort expended during online spoken discourse processing. Although inconsistent with the predictions of dual-coding theory, these results suggest that predictive coding accounts of multimodal cue integration and growth point theory can conjointly account for how beat gesture is integrated with contrast and other cues in spoken discourse.

Acknowledgments

This work was supported by a Hilibrand Postdoctoral Fellowship to L.M.M. and by award MH107426 from the National Institutes of Mental Health to J.C.M. All stimuli, analysis scripts, data, and appendices are publicly available via the Open Science
Framework at https://osf.io/fy6wp. The authors thank Simone Hasselmo and Julie Trapani for assistance with audio and video recording (respectively); Ellen Macaruso, Jake Feiler, Sarah Hughes Berheim, Cailee Nelson, Nan Mu, and Nathaniel Shannon IV for assistance with stimulus preparation and paradigm testing; Talena Day, Kathryn McNaughton, Autumn Christafore, Julia Hahn, and Alyssa Schoonmaker for assistance with data collection; and Maggie Paul for assistance with the norming study. In addition, the authors thank the attendees of the 2019 meeting of the CUNY Conference on Human Sentence Processing and the 2018 meetings of the Cognitive Science Society and the Psychonomic Society for their feedback on portions of this research.

Open Research badges

This article has earned Open Data and Open Materials badges. Data and materials are available at https://osf.io/fy6wp/.

Notes

1. Multiple systems for analyzing and transcribing English intonation exist, such as ToBI (Beckman & Elam, 1997; Silverman et al., 1992) and RaP (Breen et al., 2012; Dilley & Brown, 2005), which make somewhat different claims about the nature and type(s) of pitch accents that may exist. In the present work, we view our broad question about the effects of contrastive accenting on online reference resolution as one that generalizes across the nuances of particular intonational systems and theories.

2. Although McNeill specified that beat gesture takes the form of flicks, recent research (Prieto et al., 2018; Shattuck-Hufnagel et al., 2016) indicates that the form of beat gesture varies and that its defining feature is temporal synchrony with speech prosody.

3. To control for possible associations between the side on which the beat gesture occurred and the side on which the target object appeared, horizontally flipped duplicates were created for all videos, and presentation of original and flipped videos was counterbalanced across trials and participants by list (see Tables S4 and S5).

4. Because the original soundtracks of video clips were deleted, it was not possible to analyze the acoustic features of pitch accented words as was done for the audio-recorded spliced sentences that replaced them.

5. This occurred in <1% of trials.

6. We also considered alternate models that included target location and gesture orientation as additional control variables. For both experiments, including these
variables did not affect any of the critical results, which is to be expected because they were orthogonal to the independent variables of interest, so they were excluded in the final models.

7. For neither-contrast filler sentences, filler sentences from Experiment 1 were used as carrier sentences to create filler sentences for Experiment 2, preserving the color word + “again” combination, which sounded more natural than splicing color words and “again” together.

8. Because no motion occurred during this final video portion, extending the last frame looked relatively similar to the original videos, although the freeze frame was noticeable.

References


Supporting Information

Additional supporting information may be found online in the Supporting Information section at the end of the article:

Supplementary results.