Visuo-spatial Working Memory and the Comprehension of Iconic Gestures

Ying Choon Wu (yingchoon@gmail.com)
Institute for Neural Computation (Mail Code 0523), 9500 Gilman Drive
La Jolla, CA 92093 USA

Bonnie Chinh (bchinh@ucsd.edu)
Cognitive Science (Mail Code 0515), 9500 Gilman Drive
La Jolla, CA 92093 USA

Seana Coulson (scoulson@ucsd.edu)
Cognitive Science (Mail Code 0515), 9500 Gilman Drive
La Jolla, CA 92093 USA

Abstract

Multi-modal discourse comprehension requires speakers to combine information from speech and gestures. To date, little research has addressed the cognitive resources that underlie these processes. Here we used a dual task paradigm to test the relative importance of verbal and visuo-spatial working memory in speech-gesture comprehension. Healthy, college-aged participants encoded either a series of digits (verbal load) or a series of dot locations in a grid (visuo-spatial load), and rehearsed them (secondary memory task) as they performed a (primary) discourse comprehension task. The latter involved watching a video of a man describing household objects, viewing a picture probe, and judging whether or not the picture was related to the video. Following the discourse comprehension task, participants recalled either the verbally or visuo-spatially encoded information. Regardless of the secondary task, performance on the discourse comprehension task was better when the speaker's gestures were congruent with his speech than when they were incongruent. However, the congruency advantage was smaller when the concurrent memory task involved a visuo-spatial load than when it involved a verbal load. Results suggest that taxing the visuo-spatial working memory system reduced participants' ability to benefit from the information in congruent iconic gestures.

Keywords: depictive gesture; iconic gesture; language comprehension; multi-modal discourse; working memory

Introduction

During multi-modal discourse comprehension, listeners are tasked with integrating visual information conveyed in speakers' gestures with semantic information conveyed by their speech. Utilizing gestural information likely recruits working memory (WM) resources because it relates to linguistic information at varying levels of granularity, such as the word-, phrase, and sentence-levels (Kendon, 2004). Here we investigate the relative import of verbal versus visuo-spatial working memory resources for multi-modal discourse comprehension.

Prior research on multi-modal discourse comprehension has used a picture probe classification task in which participants view a multi-modal discourse prime followed by a picture probe that they judge as either related or unrelated to the previous stretch of discourse (Wu & Coulson, 2014). Reaction times for related picture probes are typically faster following discourse primes with *congruent* gestures that match the concurrent speech, than *incongruent* gestures that do not, suggesting congruent iconic gestures help convey information about the discourse referents (Wu & Coulson, 2014).

Consistent with the suggestion that speech-gesture integration recruits the visuo-spatial WM system, the magnitude of these congruity effects has been shown to be larger in participants with greater visuo-spatial WM capacity (Wu & Coulson, 2014). Moreover, imposing a concurrent *verbal load* during this task yielded *additive effects* of gesture congruity and WM load, while a concurrent *visuo-spatial load* yielded *interactive effects*, as gesture congruity effects were greatly attenuated under conditions of high visuo-spatial load (Wu & Coulson, 2014).

Prior research thus suggests speech-gesture integration recruits cognitive resources shared by visuo-spatial WM load tasks. One shortcoming of this earlier research, however, is that the impact of verbal load on gesture comprehension was assessed in one group of participants, while the impact of visuo-spatial load was assessed in another. Observed differences in verbal versus visuo-spatial load might reflect incidental differences in the underlying cognitive abilities of the two groups of participants, or differences in the strategies each employed.

The former possibility is particularly salient in view of models of working memory that emphasize the importance of individual differences in domain general abilities in executive function over modality specific working memory systems (e.g. Engle, 2002). According to such models, working memory capacity differences arise from domain general differences in the ability to maintain recently encoded information in the face of intervening information. Such executive attention models could explain results reported by Wu & Coulson (2014) as reflecting group differences in executive attention and fluid intelligence.

In the present study, we utilized similar picture probe classification task to Wu & Coulson (2014). However, we adopted a within- participants design to directly compare the impact of a concurrent verbal versus visuo-spatial load on

multi-modal discourse comprehension. The logic of this dual task paradigm is that if the two tasks recruit shared cognitive resources, performance of the secondary task will impair performance on the primary one. In the present study, the primary task was the picture probe classification task described above.

Participants' ability to integrate information in the speech and gestural channels was indexed in this paradigm by faster responses following congruent than incongruent gestures. Consequently, if the secondary tasks divert cognitive resources from the primary task, it would be indexed by the reduction or the elimination of congruency effects. That is, the presence of a large congruency effect, even under conditions of memory load, would suggest that the resources used in the two tasks are largely independent of one another. Alternatively, a small congruency effect would signal that resources needed for speech-gesture integration were unavailable due to the demands of the secondary memory task.

Given that the secondary tasks used here have previously been shown to be roughly matched for difficulty (Wu & Coulson, 2014), the critical question is whether congruency effects are larger when the discourse comprehension task was paired with a concurrent verbal versus visuo-spatial load task. Hypothesizing that visuo-spatial WM resources are more important for speech-gesture integration than verbal WM, we predicted the congruency effects would be smaller under conditions of visuo-spatial than verbal WM load. Executive attention models, by contrast, would predict similar sized congruency effects under both sorts of loads.

Methods

Participants

60 undergraduates (39 female) gave informed consent and received academic course credit for participation. All participants were fluent English speakers.

Materials

Materials for the Primary (Discourse Comprehension) Task were identical to those used in Wu & Coulson (2014). Discourse primes were derived from continuous video footage of spontaneous discourse centered on everyday activities, events, and objects. The speaker in the video was naïve to the experimenters' purpose and received no explicit instructions to gesture.

Short segments (2 – 8 s) were extracted in which the speaker produced both speech and gesture during his utterance. Topics varied widely, ranging from the height of a child, the angle of a spotlight, the shape of furniture, swinging a golf club, and so forth. For congruent primes, the original association between the speech and gesture was preserved. To create incongruent counterparts, audio and video portions of congruent clips were swapped such that across items, all of the same speech and gesture files were presented; however, they no longer matched in meaning. Because of the discontinuity between oro-facial movements and verbal output in incongruent items, the speaker's face was blurred in all discourse primes (congruent and incongruent). In an independent norming study using a five point Likert scale, the degree of semantic match between speech and gesture in the congruent trials was rated on average as 1.6 points higher than in the incongruent trials (3.8 (SD=.8) vs. 2.2 (SD=.7)).

Related picture probes were derived from photographs depicting objects and scenes denoted by both the spoken and gestured portions of a discourse prime (see Figure 1). Unrelated filler trials were constructed by creating new prime-probe pairings that the experimenters deemed were unrelated to one another. Related and unrelated trials were counterbalanced across four randomized lists, each containing 168 trials, and such that each picture occurred as a related probe following its associated congruent and incongruent discourse primes, and as an unrelated probe following a different pair of congruent and incongruent discourse primes. No probes or primes were repeated within any list. Verbal and visuo-spatial secondary recall tasks were evenly distributed across fifty percent of each trial type.

Secondary Recall Tasks

Each participant performed two types of secondary recall tasks. The **verbal load task** involved remembering sequences of spoken numbers. The **visuo-spatial load task** involved remembering sequences of dot locations in a two-dimensional grid. During the encoding phase of the verbal task, a series of four numbers (each ranging between one and nine) were selected pseudo-randomly, and presented via digitized audio files while a central fixation cross remained on the computer screen. For the visuo-spatial task, four dots were shown sequentially in squares selected pseudo-randomly within a 4 × 4 matrix.

After the intervening primary task, participants were prompted to recall the secondary memory sets. In the case of verbal loads, an array of randomly ordered digits from 1-9 appeared in a row in the center of the screen, and participants clicked the mouse on the numbers that they remembered hearing in the order that they were presented. In the case of visuo-spatial loads, a blank 4 × 4 grid appeared, and participants clicked the mouse in the boxes where the dots had appeared in the order that they remembered seeing them. For both types of recall, written feedback (either "Correct" or "Incorrect") was shown on the monitor for half a second after the final mouse click.

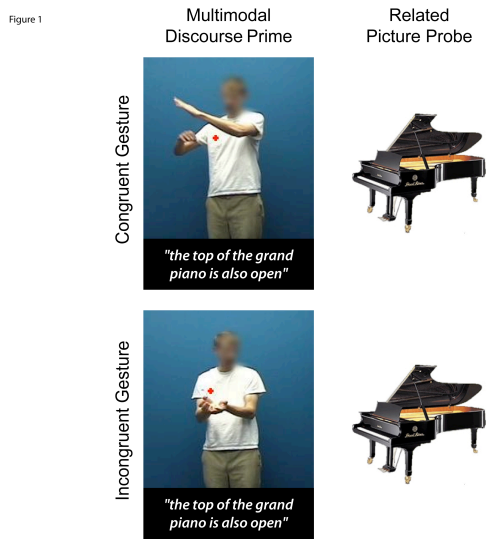

Figure 1. Primary Picture Relatedness Task

Trial Structure

As outlined in Figure 2, each trial began with a fixation cross (1s), followed by the encoding phase of the secondary task. In the case of visuo-spatial loads, each dot remained visible on the grid for one second. In the case of verbal loads, sound files lasting approximately 500ms each were presented successively with 500ms pauses in between. A half second pause concluded the encoding phase.

Primary Task The picture classification portion of each trial began with a discourse video, presented at a rate of 30ms per frame in the center of a computer monitor. A picture probe appeared in the center of the screen immediately following the video offset, and remained visible until a response was registered. Two squares labeled "Yes" versus "No" accompanied each picture at the bottom of the screen. Squares were arranged side by side, and the mouse cursor was initialized to a location equidistant between the two. Participants responded by clicking the mouse in the square labeled "Yes" on related trials and "No" on unrelated ones. No feedback was given.

Secondary Recall Task After a brief pause (250ms), participants were prompted to recall secondary memory items. Written feedback on secondary recall accuracy ("Correct" versus "Incorrect") was presented for 500ms. Between trials, the screen was blank for a half second and the mouse cursor was reset to a neutral, hidden position.

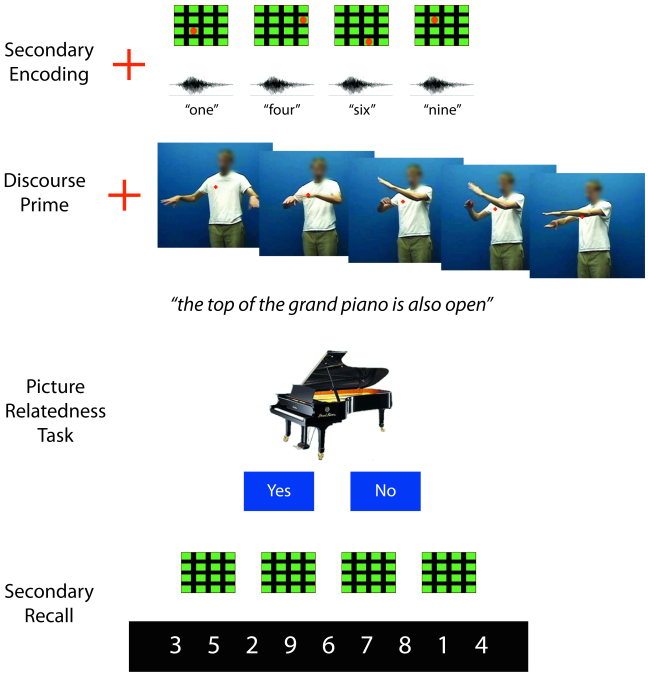

Figure 2. Trial Structure. Note that on any given trial, participants performed only one secondary memory task (encoding and recalling either the visuo-spatial load with the grid, or the verbal load with the digits).

Procedure

Participants were told they would be watching a series of short videos while rehearsing secondary memory items. Instructions began with an explanation of each kind of memory task, the verbal load task, referred to as 'Digits' trials, and the visuo-spatial task, referred to as 'Dots' trials. Participants were then told that each trial also involved a video of a man describing everyday objects and actions followed by a photograph. Participants were asked to watch and listen to each video, to respond 'yes' or 'no' whether the photograph depicted what the speaker was describing, and then to recall numbers or dot locations as prompted. Participants were encouraged to respond both as quickly and accurately as possible on the primary and secondary tasks. They were also encouraged to either visually or verbally rehearse items to be remembered. The dual-task portion of the experiment began after a short practice block comprised of two 'Dots' trials and two 'Digits' trials.

After completion of the dual-task portion of the experiment, verbal and visuo-spatial WM capacity was assessed through two short tests – the Sentence Span task (Daneman and Carpenter, 1980) and the Corsi Block task (Milner, 1971) (see Wu & Coulson, 2014 for more detail about the implementation of these measures). The Sentence Span task involved listening to sequences of unrelated sentences and remembering the sentence final word in each. All trials contained between two and five items, and were presented in blocks of three. An individual's span was the highest consecutive level at which all sentence final words

were accurately recalled (in any order) on at least two of the three trials in a block.

In the Corsi Block task, an asymmetric array of nine squares was presented on a computer monitor. On each trial, between three and nine of the squares flashed in sequence, with no square flashing more than once. Participants reproduced patterns of flashes immediately afterwards by clicking their mouse in the correct sequence of squares. An individual's Corsi span was the highest level at which at least one sequence out of five was correctly replicated (Conway et al., 2005). The entire experimental session lasted approximately two hours.

Analysis

Data from four participants were excluded due to chance level accuracy on the primary task. Response latencies were computed from the onset of the picture probe to the time of the key press. Only correct responses to related probes were analyzed, yielding a 2 (Congruent/Incongruent Discourse Primes) x 2 (Verbal/Visuo-Spatial Recall) design. Response times were trimmed within 2.5 standard deviations of each participant's mean response time. On average, 3% (sd=1%) of the data were lost. RTs by subjects, as well as proportions of accurate responses, underwent two-way repeated measures ANOVA, and follow-up contrasts were performed with t-tests. A repeated measures ANOVA test with the same factors was also conducted on recall accuracy of secondary memory items (numbers or dot locations).

Results

Secondary Recall Accuracy Rates

Secondary memory items were recalled slightly more accurately on trials involving congruent than incongruent discourse primes – 83.4% versus 81% correct (Congruency $F(1,55)=5$, $p<0.05$). Sequences of digits were recalled reliably more accurately than spatial locations of the dots -- 89% versus 79% correct (Load Modality $F(1,55)=70$, $p<0.05$). No interaction between these factors was detected ($F < 1$, n.s.).

Primary Picture Classification Accuracy Rates

On average, pictures were classified slightly more accurately when primed by congruent (93% correct, SD = 7%) than incongruent (90% correct, SD = 10%) speech and gestures (Congruency $F(1,55)=22$, $p<0.05$). No reliable effect of load modality or load modality × congruency interaction was detected ($F's < 1$, n.s.).

Primary Picture Classification Response Times

Pictures were classified more rapidly when primed by congruent than incongruent speech and gestures (Congruency $F(1,55)=43$, $p <0.05$). They were also classified more rapidly when the secondary task involved a visuo-spatial versus verbal load (Load Modality $F(1,55)=24$, $p<0.05$). Crucially, main effects were qualified by an interaction between speech-gesture congruency and load modality ($F(1,110)=5.2$, $p<0.05$). Follow-up t-tests revealed a reliable speech-gesture congruency effect when participants engaged in both visuo-spatial ($t(55)=-4$, $p<0.5$) and verbal ($t(55)=-5$, $p<0.05$) rehearsal. However, as can be seen in Figure 3, the magnitude of this effect was considerably smaller when the secondary task involved sequences of dot locations (visuo-spatial load) versus digits (verbal load).

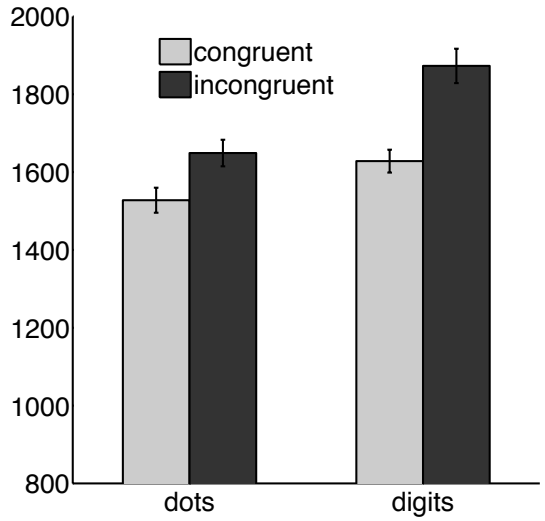

Figure 3. RTs for Discourse Comprehension task with a concurrent load on visuo-spatial (dots) versus verbal (digits) WM

Individual Differences

Finally, we modeled the relationship between WM abilities and sensitivity to gesture through two multiple regressions. The dependent variable was the magnitude of the discourse congruency on decision times under either a verbal or visuo-spatial load. Span scores on the Corsi Block and Sentence Span tasks served as predictor variables. All measures were normalized.

Multiple regression analysis indicated that the magnitude of the speech-gesture congruency effect under visuo-spatial load was reliably predicted by Corsi Block ($\beta=0.74$, $t(51)=2.3$, $p<0.05$), but not Sentence Span scores ($\beta=0.21$, $t<1$, n.s.). That is, individuals with superior visuo-spatial abilities tended to exhibit greater sensitivity to speech-gesture congruency while rehearsing dot locations. A similar analysis of the congruency effect under verbal load failed to reveal any relationship between the Corsi and Sentence Span predictor variables and participants' sensitivity to speech-gesture congruency.

Discussion

Speech-gesture congruency effects were less pronounced with the concurrent visuo-spatial than the verbal load task, consistent with our suggestion that understanding iconic

gestures recruits visuo-spatial memory resources. According to our visuo-spatial resources hypothesis, the meaning of iconic gestures is often difficult to discern until they can be mapped onto concepts evoked by the speech. The visuo-spatial WM system is used to store gestural information until it can be matched and integrated with verbally evoked concepts.

Participants' ability to benefit from the information conveyed by gestures was manifested in the present study by reliable congruency effects in all three dependent measures: recall accuracy on the secondary memory tasks, picture classification accuracy on the discourse comprehension task, and faster reaction times on the discourse comprehension task. Because there was greater overlap between the processing demands of the discourse comprehension task and the visuo-spatial load task, congruency effects were less pronounced when participants were tasked with remembering visuo-spatial information.

Because all participants performed both concurrent tasks, observed results are less amenable to explanation via domain general models of WM that emphasize the role of executive attention in these phenomena. Domain general models suggest performance on the primary task depends on participants' ability to switch fluidly between the tasks, and to suppress information that might interfere with a correct response. Assuming the secondary tasks were matched for difficulty, such models incorrectly predict similar sized congruency effects with both verbal (digits) and visuo-spatial (dots) memory loads.

Accordingly, it is critical to establish that the two concurrent load tasks placed similar demands on executive functions. In the present study, this is somewhat difficult to conclusively establish because, while reaction times on the primary task were faster with visuo-spatial than verbal loads, accuracy rates on the secondary recall tasks were greater for verbal than visuo-spatial loads. The observed speed-accuracy trade-off is consistent with our suggestion that visuo-spatial and verbal load affect speech-gesture integration processes differently, but makes it difficult to evaluate whether one task is generally more demanding than the other.

We find the suggestion that the visuo-spatial WM task was more difficult than the verbal WM task rather unlikely in view of previous work in our laboratory. Wu & Coulson (2014) used these same visuo-spatial and verbal load tasks in a dual task paradigm in which the primary task involved searching for a target letter in an array of distractors (viz., a visual search task). The visual search task has previously been used in this way to compare the demands of concurrent load tasks by evaluating how search time increases with increasing numbers of distractors, with the slope of this set-size function serving as an index of the difficulty of the secondary task (Treisman & Gelade, 1980). Critically, Wu & Coulson (2014) found similar slopes for the distractor set-size function in both concurrent tasks, suggesting they place similar demands on executive function.

Visuo-spatial WM and Iconic Gestures

Our finding that visuo-spatial WM helps mediate multi-modal discourse comprehension is consistent with existing research on speech-gesture integration. For example, Wu and Coulson (2011) describe evidence suggesting that gestures are interpreted through image-based semantic analysis – analogous to the manner whereby objects in a picture are discerned through the analysis of contours and shapes. Additionally, it has been shown that listeners use information in gestures to formulate spatially specific conceptual models of speaker meaning (Wu & Coulson, 2007). For instance, if a speaker says, "green parrot, fairly large," while indicating in gesture the bird's size and location (perched on his forearm), listeners find it easier to comprehend a pictorial depiction of a green parrot perched on a forearm relative to a green parrot in a different location, such as a cage (Wu & Coulson, 2010).

Grounded theories of language have advanced the view that mental simulations of this type are part of every day language comprehension and reasoning. Unremarkable sentences such as, "The ranger saw the eagle in the sky," have been shown to prompt faster categorization and naming of a matching picture probe depicting an eagle in flight than a mismatched probe depicting an eagle in a nest (Zwaan, Stanfield, & Yaxley, 2002), as would be expected if listeners were mentally simulating visualizable aspects of the sentence's meaning.

Likewise, in tasks such as feature generation and property verification, participants' responses have been shown to be modulated by the implied perspective from which the cue is presented (see Barsalou, 2008 for a review). For example, participants generate internal features such as *seeds* much more frequently in response to objects whose internal structure is visible (e.g. *half watermelon)* than occluded (e.g. *watermelon)* (Wu & Barsalou, 2009). When prompted to conceptualize objects from either an internal perspective (*driving a car*) or an external one (*washing a car*), adults have also been shown to categorize parts of the object more rapidly when they agree with the cued perspective (e.g. *steering wheel* and *door handle* for internal and external perspectives, respectively) (Borghi, Glenberg, and Kaschak, 2004).

In light of findings such as these, gestures may be viewed as material prompts or scaffolding that can enhance mental simulation processes regularly performed by listeners. Indeed, Hostetter & Alibali (2008) suggest that the *production* of gestures is the bodily manifestation of sensorimotor simulation processes that underlie the speakers' conceptualization of their messages. Here we suggest the comprehension of gestures also activates sensorimotor aspects of conceptual structure relevant for understanding the speaker's message.

In sum, results of the present study suggest visuo-spatial WM resources are needed to fully benefit from the information in iconic gestures. This discovery is consistent with the idea that co-speech iconic gestures promote image-based simulations of the meaning of an utterance, at least

for the descriptions of concrete objects and actions employed in the present study. Given that iconic gestures depict visual and spatial properties such as shape, size, and relative position, it is perhaps relatively unsurprising that listeners recruit visuo-spatial resources to relate information conveyed in speech to visual information conveyed in the accompanying gestures. One critical issue for future research is whether such findings extend to the gestures accompanying more abstract topics.

Acknowledgments

This work was supported by a McNair Fellowship to BC, and an NSF grant to SC (#BCS-0843946).

References

Barsalou, L. W. (2008). Grounded Cognition. *Annual Review of Psychology, 59*, 617-645.

Borghi, A. M., Glenberg, A. M., & Kaschak, M. P. (2004). Putting words in perspective. *Memory and Cognition, 32*(6), 863-873.

Daneman, M., & Carpenter, P. A. (1980). Individual differences in working memory and reading. *Journal of Verbal Learning and Verbal Behavior*, 19(4), 450–466.

Engle, R. W. (2002). Working memory capacity as executive attention. *Current directions in psychological science*, *11*(1), 19-23.

Hostetter, A. B., & Alibali, M. W. (2008). Visible embodiment: Gestures as simulated action. *Psychonomic bulletin & review*, *15*(3), 495-514.

Kendon, Adam. (2004). *Gesture: Visible action as utterance*. Cambridge University Press.

Milner, B. (1971). Interhemispheric differences in the localization of psychological processes in man. *British Medical Bulletin*, 27, 272–277.

Treisman, A. M., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive psychology*, *12*(1), 97-136.

Wu, L. L., & Barsalou, L. W. (2009). Perceptual simulation in conceptual combination: Evidence from property generation. *Acta psychologica*, *132*(2), 173-189.

Wu, Y. C., & Coulson, S. (2007). How iconic gestures enhance communication: An ERP study. *Brain and Language, 101*(3), 234-245.

Wu, Y. C., & Coulson, S. (2010). Gestures modulate speech processing early in utterances. *NeuroReport, 21*(7), 522-526.

Wu, Y. C., & Coulson, S. (2011). Are depictive gestures like pictures? Commonalities and differences in semantic processing. *Brain and language*, *119*(3), 184-195.

Wu, Y. C., & Coulson, S. (2014). Co-speech iconic gestures and visuo-spatial working memory. *Acta psychologica*, *153*, 39-50.

Zwaan, R. A., Stanfield, R. A., & Yaxley, R. H. (2002). Language comprehenders mentally represent the shapes of objects. *Psychological science*, *13*(2), 168-171.