Working Memory and Co-Speech Iconic Gestures

Seana Coulson (scoulson@ucsd.edu) and Ying Choon Wu (yingchoon@gmail.com)

Department of Cognitive Science

9500 Gilman Drive, La Jolla 92093 USA

Abstract

The importance of verbal and visuospatial working memory (WM) for co-speech gesture comprehension was tested in two experiments using the dual task paradigm. Healthy, college-aged participants encoded either a dot locations in a grid (Experiment 1), or a series of digits (Experiment 2), and rehearsed them as they performed a discourse comprehension task. The discourse comprehension task involved watching a video of a man describing household objects, and judging which of two words probes was most related to the video. Following the discourse comprehension task, participants recalled either the verbally or visuo-spatially encoded information. In both experiments, performance on the discourse comprehension task was faster when gestural information was congruent with the speech than when it was incongruent. Moreover, performance on the discourse comprehension task was impacted both by increasing the load on the visuospatial WM system (Experiment 1) and the verbal WM system (Experiment 2). However, in both studies effects of WM load and gesture congruency were additive, suggesting they were independent.

Keywords: depictive gesture; discourse comprehension; iconic gesture; multimodal meaning; representational gesture; verbal working memory; visuospatial working memory

Introduction

Co-speech gestures, which are produced spontaneously in co-ordination with speaking, offer an exciting opportunity to explore the relationship between body movement and higher order cognitive functions, such as language comprehension and conceptualization (for review, see Goldin-Meadow, 2003). To date, little research has addressed the cognitive resources that allow us to understand these gestures and to relate their meaning to that conveyed by the accompanying speech. Because gestures relate to linguistic information at varying levels of granularity, including the word-, phrase, and sentence- levels (Kendon, 2004), one fairly straightforward possibility is that working memory (WM) plays an important role in these processes, allowing listeners to maintain information conveyed in the gestural stream until it can be integrated with relevant information presented in the speech.

Previous research has contrasted the *verbal resources hypothesis*, that speech gesture integration primarily recruits verbal WM, with the *visuo-spatial resources hypothesis*, that speech gesture integration recruits the visuo-spatial WM system. That work employed a discourse comprehension task in which participants viewed a multi-modal discourse prime of a speaker describing everyday objects, followed by a picture that participants judged as either related or unrelated to the prime (Wu &Coulson, 2014). Reaction times for related picture probes are typically faster following discourse primes with congruent gestures that match the concurrent speech, relative to incongruent gestures that do not, suggesting congruent iconic gestures help convey information about the discourse referents (Wu & Coulson, 2014).

Consistent with the visuo-spatial resources hypothesis, 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 that speech-gesture integration recruits cognitive resources shared by visuo-spatial WM load tasks.

One shortcoming of research by Wu and Coulson (2014) is that their measure of speech-gesture integration involved participants' responses to picture probes that followed videos of multimodal discourse. Given that responding to pictorial stimuli presumably imposes a load on participants' visuospatial processing resources, this task may overestimate the importance of visuospatial WM for the comprehension of co-speech gestures.

The present study explored the role of verbal versus visuospatial WM in speech-gesture integration by utilizing a dual task paradigm similar to that in Wu & Coulson (2014). However, rather than using performance on a picture probe task to index comprehension of the gestures, we asked participants to choose which of two words was most related to the preceding discourse video. Experiment 1 paired this

discourse comprehension task with a visuospatial WM task, and Experiment 2 paired it with a verbal WM task.

Experiment 1

Experiment 1 tests how increasing the load on participants' visuospatial WM system impacts their sensitivity to the meaning of co-speech gestures in multimodal discourse. The logic of the dual task paradigm is that if the two tasks recruit shared cognitive resources, engagement in the secondary task will impair performance on the primary one. In Experiment 1, the primary task is that of discourse comprehension, as indexed by a word probe task, while the secondary task involved memory for a sequence of dot locations in a grid. We manipulated the difficulty of multi-modal discourse comprehension by varying the semantic congruity of the gestures and the speech in our discourse videos. The difficulty of the visuospatial recall task was varied by asking participants to remember a sequence of either four locations (high load), or to remember a single location (low load). Consequently, if the recall task diverts cognitive resources from speech-gesture integration, it would be reflected in a change in the congruency effects as a function of visuospatial load – that is, either the amplification of congruency effects, the reduction of congruency effects, or their elimination altogether.

Methods

Participants Participants were 51 healthy undergraduates who, in exchange for participation, received extra credit for a course in cognitive science, linguistics, or psychology.

Materials A total of 84 discourse primes were kindly provided by Dr. Wu. These 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 (2-8s) 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 all items, all of the same speech and gesture files were presented; however, they no longer matched in meaning.

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 (congruent = 3.8, sd=0.8 versus incongruent = 2.2, sd = 0.7). Because of the discontinuity between oro-facial movements and verbal output in incongruent items, the speaker's face was blurred in all discourse primes (i.e. both the congruent and incongruent version of each).

Each discourse prime was followed by the presentation of two word probes arrayed vertically in the center of the monitor. The *related* probe was a word related to the audio content of the video, and was intended to specifically highlight the semantic content of the congruent gesture. The *unrelated* probe was intended to be unrelated to any aspect of the audio or video. The same two word probes followed the congruent and the incongruent version of each audio file. The location of the related probe (i.e. at the top or the bottom of the array) was chosen randomly on each trial.

Half of the trials (n=42) were accompanied by a low load version of the visuo-spatial recall task, and half with a high load version of the same task. The visuospatial recall task was similar to the dot movement task employed by Wu & Coulson (2014), in which participants were asked to remember a single location in a 4 x 4 grid on low load trials, and an ordered sequence of four locations on high load trials. The gesture congruity and memory load manipulations were fully counterbalanced.

Procedure Each trial began with a fixation cross (1s), followed by the encoding phase of the secondary task (visuospatial WM). Secondary encoding involved the visual presentation of a sequence of dots in a 4x4 grid. High load trials involved a sequence of four distinct locations, while low load trials involved the presentation of a single dot. Each dot remained visible on the grid for one second. A 500ms pause concluded the encoding phase.

The discourse comprehension portion of each trial began with a discourse video, presented at a rate of 30ms per frame in the center of a computer monitor. Immediately following the video offset, the probes appeared above and below the fixation cross. The mouse cursor was initialized to a location equidistant between the two. Participants were asked to respond by clicking the mouse in the square that contained the word that best matched the scenario described by the speaker. No feedback was given.

After a 250ms pause, participants were prompted to recall the location of dots in the grid in the order that they had been presented. Written feedback ("correct" versus "incorrect") was provided following each trial for 500ms. Between trials, the screen was blank for half a second and the mouse cursor was reset to a neutral hidden position.

After completion of the dual-task portion of the experiment, verbal and visuo-spatial WM capacity were assessed through two short tests – an auditory version of the Sentence Span task (Daneman and Carpenter, 1980) and a computerized version of the Corsi Block task (Milner, 1971). The Listening 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.

Results

Visuospatial Recall Task Performance on the visuo-spatial recall task was indexed by the number of trials in which the participant correctly recalled all of the to be remembered locations. These values were subjected to repeated measures ANOVA with factors memory load (High, Low) and gesture congruity (Congruent, Incongruent). This analysis revealed only an effect of Load, $F(1, 67) = 130.2$, $p < 0.05$, $ges = .24$. Figure 1 shows the average number of correct trials in each condition and clearly indicates better performance in trials with a low load (1 dot location) than in the high load trials (4 dot locations). These data suggest the memory load task was more difficult in the high than the low load condition.

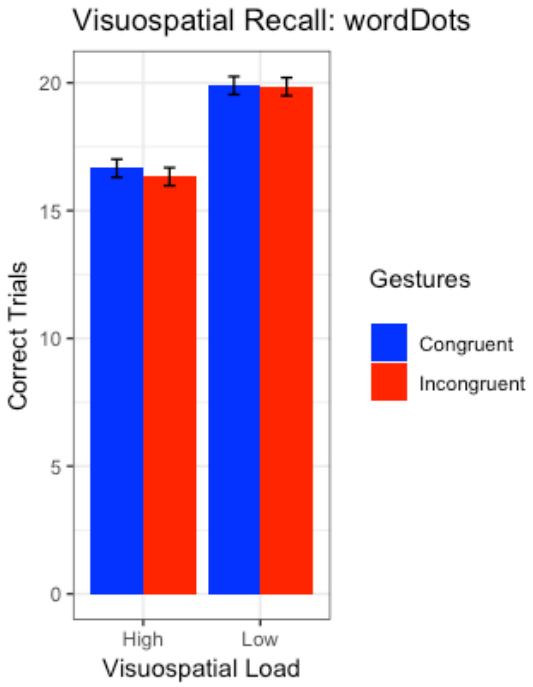

Figure 1: Mean number of correct trials on the recall task in each condition of Experiment 1.

To test the importance of WM capacity for sensitivity to our memory load manipulation, we computed the difference between each participant's accuracy on the high and low load trials. We then constructed a linear model to predict this difference due to the load manipulation as a function of participants' scores on the Corsi Block and Listening Span tasks. This model significantly predicted accuracy on the recall task, $F (2, 64) = 5.18$, $p < 0.01$, accounting for 13.95% of the variance. ANOVA on the output of the model suggested scores on the Corsi Block Task served as significant predictors, $F (1, 64) = 10.3$, $p < 0.01$, while scores on the Listening Span did not, $F (1, 64) = 0.03$, n.s. The systematic relationship between scores on the Corsi Block Task with the visuo-spatial load effect supports our contention that the dots task recruits visuo-spatial WM.

Discourse Comprehension Task

Accuracy on the discourse comprehension task was scored by counting the number of correct trials in each condition for each participant. Figure 2 shows the mean scores in each condition. These values were subjected to repeated measures ANOVA with factors gesture congruity (congruent/incongruent) and memory load (high/low). This analysis revealed a main effect of gesture congruity, $F(1, 66) = 3.4$, $p < 0.05$, ges 0.08, as participants were more accurate when speech was accompanied by congruent than incongruent gestures. Memory load was not significant, either as a main effect, $F(1, 66) = 2.5$, n.s., or as an interaction with gesture congruity, $F(1, 66) = 1.08$, n.s.

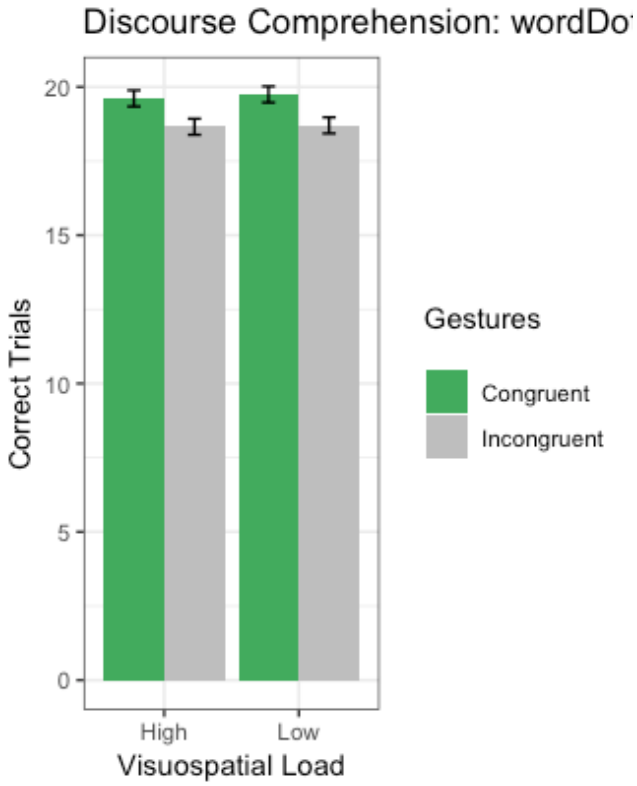

Figure 2: Mean number of correct trials in the discourse comprehension task in Experiment 1. Error bars represent 95% confidence intervals.

To explore the relationship between sensitivity to gestures and our individual difference measures, we computed the difference between the total number of trials each participant responded to correctly in the congruent gesture condition and the incongruent gesture condition. A linear model was constructed to predict this difference score from the Corsi Span score and the Listening Span score. This model accounted for 10.4% of the variance in difference scores, $F(2, 64) = 3.72$, $p < 0.05$. ANOVA on the output of

the model suggested only Corsi Span scores served as a significant predictor, $F(1, 64) = 5.3$, $p < 0.05$.

Response times for correct trials on the discourse comprehension task were analyzed with linear mixed effects models with fixed effects for gesture congruity and visuospatial load, and random effects for subject and item (viz., the audio file held constant across congruent and incongruent gesture versions of each stimulus). Random effect structure was determined via backwards model comparison using the step function in lmerTest, beginning with the 'maximal' structure allowed by the design.

Mean response times in each condition are shown in Figure 3. Performance on this task was an additive function of gesture congruity, $t = -6.84$, $p < 0.001$, with responses that were on average 383ms faster following congruent than incongruent gestures, and memory load, $t = -3.95$, $p < 0.001$, with responses an average of 170ms faster in high load trials than low load trials. The latter presumably results because participants desire to rush through the discourse comprehension task in order to 'unload' memory items in the recall task that immediately followed.

Discussion

Experiment 1 suggests a relationship between visuospatial WM capacity and sensitivity to speech-gesture congruity, but fails to support a causal link between visuospatial WM and the comprehension of gestures.

First, did the visuospatial recall task (viz. the dot task) serve to divert visuospatial resources from the primary task? Indeed, recall performance was worse under conditions of high than low load. Moreover, participants' performance on the dot task was systematically related to their visuospatial WM capacity as indexed by their scores on the Corsi block task. These data suggest that the dot task did indeed recruit our participants' visuospatial processing resources, thereby making them less available for primary task performance.

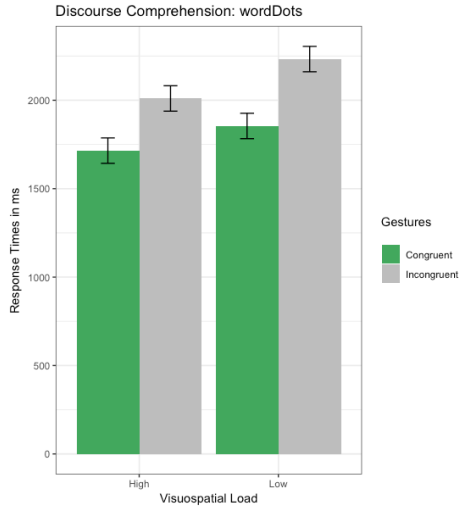

Figure 3: Mean response times in the discourse comprehension task in Experiment 1. Error bars depict 95% confidence intervals.

Second, was the discourse comprehension task employed here sensitive to the relationship between the gestures and the speech? Participants responded more quickly and more accurately on trials preceded by discourse with congruent gestures than incongruent ones. Experiment 1 thus extends results reported in Wu & Coulson (2014), showing that the facilitative impact of congruent gestures can be detected with the word probe paradigm employed in the present study. Moreover, as in the report by Wu and Coulson (2014), the participants who scored the highest on our independent assessment of visuospatial WM capacity were those who showed the largest gesture congruity effects.

Finally, how was performance of the discourse comprehension task impacted by the diversion of visuospatial processing resources? Apart from the gesture congruity effect noted above, the discourse comprehension task was also impacted by visuospatial load. Load had a somewhat paradoxical impact on responses as participants responded faster but less accurately on high load trials. Importantly, though, these two effects were additive, suggesting the discourse comprehension task proceeded somewhat independently of the visuo-spatial recall task.

Experiment 2

Experiment 2 paired the discourse comprehension task with a verbal WM task to explore how reducing the availability of verbal resources impacted participants' sensitivity to iconic co-speech gestures.

Methods

Audio and video materials were identical to those used in Experiment 1, as were the word probes. As in Experiment 1, half of the trials were accompanied by a low load recall task, and half with a high load recall task. The secondary recall task was similar to the digit recall task employed by Wu & Coulson, in which participants were asked to remember a single digit on low load trials, and an ordered series of four digits on high load trials. As in Experiment 1, the gesture congruity and memory load manipulations were fully counterbalanced.

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. As for the visuospatial WM task in Experiment 1, the stimulus onset asynchrony for to-be-remembered items was 1 second.

During the recall phase of the task, 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. Written feedback (either "Correct" or "Incorrect") on the recall task was shown on the monitor for half a second after the final mouse click.

Results and Discussion

Verbal Recall

Performance on the verbal recall task was indexed by the number of trials in which the participant correctly recalled all of the to be remembered digits (see Figure 4). These values were subjected to repeated measures ANOVA with factors memory load (High/Low) and gesture congruity (Congruent/Incongruent). This analysis revealed only an effect of memory load, $F(1, 47) = 35.9$, $p < 0.05$, $ges = .13$. Figure 4 shows the average number of correct trials in each condition and clearly indicates better performance in the low load (1 digit) trials than in the high load trials (4 digits). These data suggest the task worked as intended to occupy verbal WM.

To test the importance of WM capacity for sensitivity to our verbal memory load manipulation, we computed the difference between each participant's accuracy on the high and low load trials. We then constructed a linear model to predict this memory load effect as a function of scores on the Corsi Block and Listening Span tasks. This initial model only approached significance, $F (2, 45) = 2.92$, $p = 0.06$. Backwards model selection via the step function in the MASS package in R indicated that the best model of memory load effects was one that included a single factor, participants' Listening Span scores.

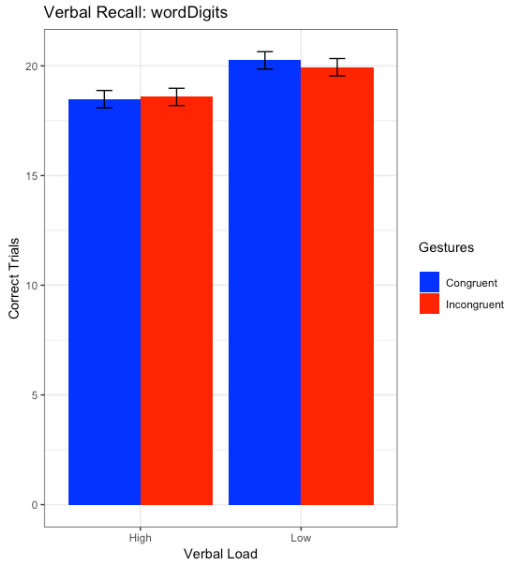

Figure 4: Mean number of correct trials in the verbal recall task for each condition in Experiment 2. Error bars represent 95% confidence intervals.

Accordingly, Corsi Span scores were dropped. The reduced model was significant, $F(1, 46) = 5.81$, $p < 0.05$, predicting 11.2% of the variance. The coefficient on the Listening Span factor was -1.2, indicating the load effect was most pronounced in participants with the lowest Listening Span scores. These data indicate a relationship between sensitivity to the digit load manipulation with our independent assessments of participants' verbal WM capacity, consistent with our assumption that the digit recall task diverted verbal WM resources. The systematic relationship between scores on the Listening Span Task with the verbal load effect supports our contention that the digit recall task recruits verbal WM resources.

Discourse Comprehension

Accuracy on the discourse comprehension task was scored by counting the number of correct trials in each condition for each participant. Figure 5 shows the mean scores in each condition. These values were subjected to repeated measures ANOVA with factors gesture congruity (congruent/incongruent) and memory load (high/low). This analysis revealed a main effect of gesture congruity, $F (1, 47) = 3.4$, $p < 0.05$, ges 0.12, as participants were more accurate when speech was accompanied by congruent than incongruent gestures. Memory load was not significant, either as a main effect, $F (1, 47) = 0.03$, $ges < 0.01$ or as an interaction with gesture congruity, $F (1, 47) = 1.34$, n.s, $ges < 0.01$.

To explore the relationship between sensitivity to gestures and our individual difference measures, we computed the difference between the total number of trials each participant responded to correctly in the congruent gesture condition and the incongruent gesture condition. A linear model was constructed to predict this difference measure from the Corsi Span score and the Listening Span score. However, neither this model nor any of the models explored with backwards model selection provided a significant account of these effects, indicating the absence of a systematic relationship between working memory capacity and this measure of sensitivity to gesture congruity.

Response times were analyzed in the same manner as in Experiment 1. Analysis involved the construction of linear mixed effects models with fixed effects of memory load and gesture congruity, and random effects of subject and item. As in Experiment 1, backwards model selection was used to simplify the random effects structure and choose the best model.

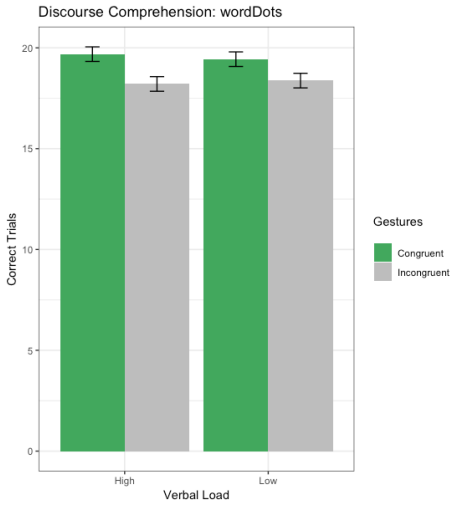

Figure 5. Number of correct trials on the discourse comprehension task in each condition of Experiment 2. Error bars depict 95% confidence intervals.

As in Experiment 1, the memory load effect results due to a 122ms faster responses in the high load trials than in the low load ones, $t = -2.55$, $p < 0.05$. Further, responses were 327ms faster in the congruent trials than the incongruent ones, $t = -5.21$, $p < 0.001$. Figure shows mean response times in each condition.

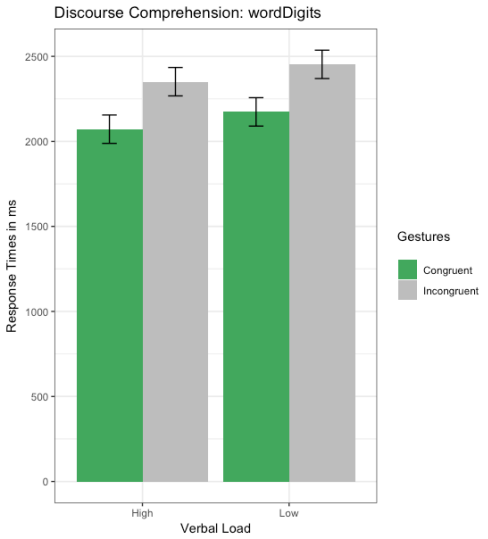

Figure 5: Mean response times for the discourse comprehension task in each condition of Experiment 2. Error bars represent 95% confidence intervals.

General Discussion

Results of the present study provide only modest support for the visuospatial resources hypothesis, and no support for the verbal resources hypothesis. Reducing the availability of visuospatial resources impacted multimodal discourse comprehension, but did not modulate participants' sensitivity to the semantic congruity of co-speech gestures. Likewise, reducing the availability of verbal resources impacted the discourse comprehension task, but did not modulate participants' sensitivity to the semantic congruity of co-speech gestures. In Experiment 1, however, sensitivity to gesture congruity was systematically greater among participants with the greatest visuospatial WM capacity. Thus, while we find no support for a direct causal role of visuospatial WM and speech-gesture integration, visuospatial resources may be relevant to some aspect of gestural processing.

Results of the present study stand in stark contrast to those reported in Coulson & Wu (2014) using the same discourse materials, the same secondary memory task, but that utilized a picture probe to test gesture comprehension rather than the word probes employed here. In tests with picture probes, Coulson & Wu (2014) found that participants were less sensitive to gesture congruity when visuospatial resources were taxed. In the present study, responses to word probes were significantly impacted by gesture congruity, but sensitivity to gestural information was similar under conditions of high and low visuospatial load. This discrepancy might result because the discourse comprehension task in Wu & Coulson (2014) was more taxing than that in the present study. Alternatively, it might be more related to the extent that the picture probe task draws more on the visuospatial resources shared with gesture processing than does the word probe task.

Indeed, the latter interpretation is consistent with the similarity between the impact of verbal memory load in Experiment 2 of the present study with that in the parallel study in Wu & Coulson (2014). Using a picture probe to assess discourse comprehension, they found that performance was impacted both by gesture congruity and by verbal memory load, although the two factors did not interact. Similarly, here we find that performance on the word probe task was independently influenced by gesture congruity and by verbal memory load. The similar impact of verbal versus visuospatial memory load on discourse comprehension as assessed with the word probes employed here also mitigates the concern raised by Wu & Coulson (2014) that the two secondary tasks differ in their demands on central processing resources.

We suggest that the greater impact of the dots task on the processing of picture probes than word probes may be indicative of the role that iconic co-speech gestures play in communication. Congruency effects on the word probe task suggest that speakers readily exploit the information in gestures to detect semantic relationships between novel words and the extant discourse context. However, perhaps because gestures are habitually used to interpret words, this process exerted minimal enough cognitive demands as to resist interference from concurrent demands on either verbal or visuospatial memory systems. By contrast, the picture probe task used by Wu & Coulson (2014) suggested that visuospatial resources were particularly important for detecting a relationship between the pictures and multimodal discourse about concrete topics.

Future research should increase the demands of either the discourse comprehension task or those of the secondary memory tasks in order to elucidate the reason for our failure to observe a differential impact of memory load on sensitivity to gestures. Perhaps titrating memory load demands individually (as in Frank, et al., 2012) will allow us to better estimate its impact on discourse comprehension.

References

Frank, M. C., Fedorenko, E., Lai, P., Saxe, R., & Gibson, E. (2012). Verbal interference suppresses exact numerical representation. *Cognitive psychology*, *64*(1-2), 74-92.

Goldin-Meadow, S., Nusbaum, H., Kelly, S. D., & Wagner, S. (2001). Explaining math: Gesturing lightens the load. *Psychological Science*, *12*(6), 516-522.

Kendon, Adam. (2004). *Gesture: Visible Action as Utterance*. Cambridge University Press.

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