Acta Psychologica 153 (2014) 39–50

Contents lists available at ScienceDirect

Acta Psychologica

journal homepage: www.elsevier.com/ locate/actpsy

Co-speech iconic gestures and visuo-spatial working memory

Ying Choon Wu [a,b,*], Seana Coulson [a,c]

[a] Center for Research in Language, UC San Diego 0526, 9500 Gilman Dr., La Jolla, CA 92093, USA
[b] Swartz Center for Computational Neuroscience, UC San Diego 0559, 9500 Gilman Dr., La Jolla, CA 92093, USA
[c] UC San Diego, Dept. of Cognitive Science 0515, 9500 Gilman Dr., La Jolla, CA 92093, USA

ARTICLE INFO

Article history:
Received 4 September 2013
Received in revised form 2 August 2014
Accepted 8 September 2014
Available online 1 October 2014

PsycINFO classification:
2343 Learning & Memory
2720 Linguistics & Language & Speech
2320 Sensory Perception

Keywords:
Dual task methodology
Gesture comprehension
Iconic gestures
Multi-modal discourse
Spatial cognition
Working memory

ABSTRACT

Three experiments tested the role of verbal versus visuo-spatial working memory in the comprehension of co-speech iconic gestures. In Experiment 1, participants viewed *congruent* discourse primes in which the speaker's gestures matched the information conveyed by his speech, and *incongruent* ones in which the semantic content of the speaker's gestures diverged from that in his speech. Discourse primes were followed by picture probes that participants judged as being either *related* or *unrelated* to the preceding clip. Performance on this picture probe classification task was faster and more accurate after congruent than incongruent discourse primes. The effect of discourse congruency on response times was linearly related to measures of visuo-spatial, but not verbal, working memory capacity, as participants with greater visuo-spatial WM capacity benefited more from congruent gestures. In Experiments 2 and 3, participants performed the same picture probe classification task under conditions of high and low loads on concurrent visuo-spatial (Experiment 2) and verbal (Experiment 3) memory tasks. Effects of discourse congruency and verbal WM load were additive, while effects of discourse congruency and visuo-spatial WM load were interactive. Results suggest that congruent co-speech gestures facilitate multi-modal language comprehension, and indicate an important role for visuo-spatial WM in these speech–gesture integration processes.

1. Introduction

Successful communication often requires multi-modal integration, whereby interlocutors combine information from the verbal channel with visual information about the speaker and the environment. For example, we have documented a speaker uttering the phrase, "manual adjustment lens," to describe a camera while making hand movements that resemble the act of focusing a telephoto lens. The speech and the gesture in this example provide complementary information — and by combining their meanings, it becomes evident that the speaker is describing the lens of a *camera*, and not some other optical device, such as a telescope or a pair of binoculars (Wu & Coulson, 2007). Although prior research indicates that listeners rapidly combine the meaning of speech and iconic gestures in examples such as this (Kelly, Kravitz, & Hopkins, 2004; Ozyurek, Willems, Kita, & Hagoort, 2007; Wu & Coulson, 2010), little is known about the cognitive resources mediating these integration processes.

Here, we focus on *depictive* or *iconic gestures* – that is, those which bear featural similarities to the concepts they represent – as prior research suggests iconic gestures impact semantic aspects of real-time discourse comprehension (Kelly et al., 2004; Ozyurek et al., 2007; Wu & Coulson, 2010). Given that iconic gestures depict visual properties such as shape and size, one obvious possibility is that visuo-spatial processes are important for listeners' success at relating information conveyed in the verbal modality to visual information conveyed in their accompanying gestures. The *visuo-spatial resources hypothesis* is a natural fit with gesture production models suggesting that people gesture in order to convey analogue information in mental images (McNeill, 1992), or to coordinate spatial aspects of a message with the propositional content in their speech (Kita, 2000). Indeed, the gesture production literature suggests that people are more likely to gesture when their speech has spatial or imagistic content (Hadar & Krauss, 1999; Hostetter & Hopkins, 2002; Lavergne & Kimura, 1987; Morsella & Krauss, 2004). Although the *comprehension* of gestures has received much less attention, the visuo-spatial resources hypothesis is consistent with research demonstrating similarities between patterns of brain response to iconic gestures and photographs of real world objects (Wu & Coulson, 2011), as well the finding that listeners use information available in speaker's iconic gestures to help formulate visually specific situation models (Wu & Coulson, 2007, 2010).

However, as their name suggests, co-speech gestures occur almost exclusively in the context of speech — and hence, their semantic analysis may depend heavily on verbal resources. The *verbal resources hypothesis* is in keeping with research suggesting that the meaning

* Corresponding author at: Center for Research in Language, UC San Diego, 9500 Gilman Drive 0526, La Jolla, CA 92093-0526, USA. Tel.: +1 858 869 7014; fax: +1 858 822 5097.
E-mail addresses: ywu@cogsci.ucsd.edu, yingchoon@gmail.com (Y.C. Wu).

http://dx.doi.org/10.1016/j.actpsy.2014.09.002
0001-6918/

of iconic gestures is highly ambiguous, and is determined largely by the meaning of the speech that accompanies them (Hadar & Pinchas-Zamir, 2004; Krauss, Dushay, Chen, & Rauscher, 1995). It is also consistent with neuroimaging research that indicates many of the brain areas mediating the interpretation of gesture, also mediate the interpretation of speech (Straube, Green, Weis, & Kircher, 2012; Willems, Ozyurek, & Hagoort, 2007). Finally, the two hypotheses are not mutually exclusive, as it is quite possible that speech–gesture integration recruits both verbal and visuo-spatial resources.

Given the function of co-speech gestures in real-time language comprehension, working memory (WM) is likely to play an important role in their interpretation. According to the now classic model advanced by Baddeley and Hitch (1974), WM is critical for online processing, serving to temporarily maintain and store perceptual information, and enabling the appropriate updating of representations in long term memory. Notably, WM is widely thought to be comprised of a central controller as well as at least two distinct, modality-specific subsystems dedicated to the maintenance of *visual* information via the visuo-spatial sketch pad, and auditory and *verbal* information via the phonological loop. If listeners tend to preferentially recruit visuo-spatial or verbal resources during speech–gesture integration, we would expect to observe a relationship between the impact of iconic gestures on discourse comprehension and the availability of either visuo-spatial or verbal WM resources (or both).

The present study explored this hypothesis using a two-fold approach. Experiment 1 adopted a correlational method, examining whether there was a relationship between individual differences in measures of either verbal or visuo-spatial WM capacity and individual differences in sensitivity to iconic gestures. In Experiments 2 and 3, we used a dual task paradigm to examine whether taxing different components of WM impact gesture comprehension, suggestive of a causal role for WM in speech–gesture integration. Accordingly, these studies assessed whether participants' ability to utilize the information in co-speech gestures was compromised by manipulating the load on either visuo-spatial (Experiment 2) or verbal (Experiment 3) WM. Finally, Experiment 4 was conducted to ensure that differences in the results of Experiments 2 and 3 did not stem from differences in the difficulty of the secondary verbal and visuo-spatial recall tasks used in those studies.

2 . Experiment 1

To explore the cognitive resources mediating speech–gesture integration, Experiment 1 examined the relationship between individual differences in WM capacity, as measured through verbal and visuo-spatial span tests, and sensitivity to speech–gesture congruency, as measured through a picture probe classification task. Healthy adults viewed short video clips of spontaneous discourse involving iconic gestures, and then classified subsequent photographs of objects and scenes (picture probes) as either related or unrelated to the discourse primes. Control primes were created by swapping the audio and video portions of the original video clips so that the gestures no longer exhibited a semantic relationship to the content of the utterance. Manipulating both the congruency of the discourse prime and the relatedness of the picture probe, we predicted that discourse congruency would have a much greater impact on related than unrelated picture probes.

Our reasoning stemmed from the following premises. First, if related picture probes were classified faster than unrelated ones, irrespective of the preceding discourse prime (a simple relatedness main effect), it could be argued that participants based their responses on the spoken aspects of each utterance, while ignoring the co-occurring gestures. Alternatively, if all probes were classified faster when primed by congruent versus incongruent speech and gestures (a simple main effect of congruency), the difference in classification times could be attributed to perceptual rather than semantic differences between the congruent and incongruent primes — that is, it could be argued that the observed effect was driven primarily by the different types of video editing performed in the construction of the discourse primes as opposed to the degree of semantic coherence between speech and gestures in the two types of trials. On the other hand, if participants were truly sensitive to speech-gesture congruency, we would expect to see much larger congruency effects in response to *related* picture probes, where congruent gestures might aid task performance, than for *unrelated* ones, where neither congruent nor incongruent gestures would be helpful.

Experiment 1 examined the relationship between the impact of speech–gesture congruency on picture probe comprehension and individual differences in verbal and visuo-spatial WM capacity. If visual analysis figured prominently in participants' abilities to conceptually integrate the speaker's speech and gestures, we would expect increasingly superior picture probe comprehension across individuals with increasing visuo-spatial WM capacity. Such an outcome would be likely if space plays a crucial role in the interpretation of iconic gestures – that is, if comprehenders draw mappings between spatially instantiated features of a gesture – such as hand shape, hand location, trajectory, rate of motion, and so forth – and imagistic representations stored in long term memory.

On the other hand, it is possible that verbal processes are more important than visuo-spatial ones for understanding multi-modal discourse. Gestures may, for example, directly activate language-based representations that impact comprehension. Alternatively, increased efficiency at verbal processing may free up attentional resources that allow for greater sensitivity to the semantic properties of gestures. In either case, the *verbal resources hypothesis* predicts a positive relationship between individuals' verbal WM capacity and the impact of iconic gestures on discourse comprehension.

Experiment 1 thus explored the verbal and visuo-spatial resources hypotheses by testing for a relationship between individual differences in verbal and visuo-spatial WM capacity and differences in sensitivity to speech–gesture congruency. Verbal WM capacity was assessed with the Sentence Span test (Daneman & Carpenter, 1980), in which participants hear progressively longer lists of unrelated sentences and recall the sentence final words when cued. Visuo-spatial WM capacity was assessed using the Corsi Block task (Berch, 1998; Milner, 1971), which involves remembering and reproducing progressively longer sequences of block locations. Through multiple regression analysis, the magnitude of the speech–gesture congruency effect on discourse comprehension was modeled using the magnitude of both verbal and visuo-spatial WM capacity as predictor variables. A positive linear relationship between verbal WM capacity and our measure of gesture sensitivity would support the verbal resources hypothesis. Likewise, a positive predictive relationship between visuo-spatial WM capacity and our measure of gesture sensitivity would support the visuo-spatial resources hypothesis.

3 . Methods

3.1. Participants

64 UCSD undergraduates (38 female) gave informed consent and received academic course credit for participation. All participants were fluent English speakers.

3.2. Span tasks

3.2.1. Corsi block task

The Corsi block-tapping task (Milner, 1971) is a widely used test of spatial skills and non-verbal WM. In the computerized variant implemented here, an asymmetric array of nine squares was presented on a monitor. On each trial, some or all of the squares would flash in sequence, though no square flashed more than once. Participants were instructed to reproduce each flash sequence immediately afterwards by clicking their mouse in the appropriate squares in the order that

the flashes had occurred. Sequences ranged from four to nine flashes and were presented in blocks of five. Successfully reproducing at least one sequence in a block led to advancement to the next level. The task terminated when no sequences were correctly reproduced, or when level nine was completed (Berch, 1998). An individual's block span was the highest level at which at least one sequence was correctly replicated (Conway et al., 2005).

3.2.2. Sentence span task

Modeled after the classic work of Daneman and Carpenter (1980), this task required that participants listen to sequences of unrelated sentences and remember the sentence final word in each. All trials contained between two and five sentences, and concluded with a cue to write down the remembered words in any order. Trials were grouped in blocks presented in order of increasing difficulty (i.e. beginning with two sentences per trial, and ending with five). Filler trials also included comprehension questions, intended to discourage participants from ignoring the meaning of the sentences, and attending only to the sentence final words. Fillers were interspersed with experimental trials, and participants were unaware that their recall on these trials was not scored. An individual's *sentence span* was the highest consecutive level at which all sentence final words were accurately recalled on at least two of the three trials in a block. Additional half points were added in cases in which a participant correctly completed at least two thirds of a block after failing to reach criterion at a lower level (St. George, Mannes, & Hoffmant, 1997; Waters & Caplan, 1996).

3.3. Picture relatedness task

3.3.1. Materials

Discourse primes were constructed from continuous video footage of a naive speaker describing everyday activities, events, and objects to an off-camera interlocutor. No explicit instructions to gesture or explanation of the experimenter's true motivation for filming were given.

Short clips (2–8 s) containing speech and gestures deemed as depictive by the researchers were extracted. Gestures represented a broad range of semantic content, including the height of a child, the angle of a spotlight, the shape of furniture, swinging a golf club, and so forth. Incongruent counterparts involved the same discourse materials. However, audio and video portions of each clip were swapped with other clips so that the degree of semantic congruence between speech and gestures was diminished (Fig. 1). Because the incongruent video primes introduced discontinuity between oro-facial movements and verbal output, the speaker's face was blurred in all discourse primes.

Ten naïve individuals rated materials for the degree of correspondence between speech and gestures on a five point Likert scale (1 = highly incongruent; 5 = highly congruent). From the original corpus, a final set of 140 items were selected with the goal of maximizing the difference between the mean rating for the congruent speech–gesture pairing relative to its incongruent counterpart (congruent minus incongruent). The average rating was 2.2 (SD = .7) for incongruent videos, 3.8 (SD = .8) for congruent ones.

Related picture probes were created from photographs agreeing with both the spoken and gestured portions of each utterance. Unrelated trials were constructed by pairing the same pictures with different discourse primes so that probes agreed with neither the speech nor gestures that immediately preceded them (see Fig. 1).

Across four randomized lists, each containing 140 trials, discourse primes and picture probes were distributed such that each stimulus served as its own control. No items were repeated within any one list, but across the four lists, each picture served 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. The camera in Fig. 1, for instance, was presented in the following four different ways over four separate recording sessions: as a related probe following the utterance, "manual adjustment lens," paired with a congruent gesture; as a related probe following the same speech paired with an incongruent gesture, as an unrelated probe following the utterance, "um – it's a corner, corner couch," paired with either a congruent or incongruent gesture. On each list, there were 35 items in each cell of the 2 (Congruent/Incongruent Discourse Primes) × 2 (Related/Unrelated Picture Probes) design.

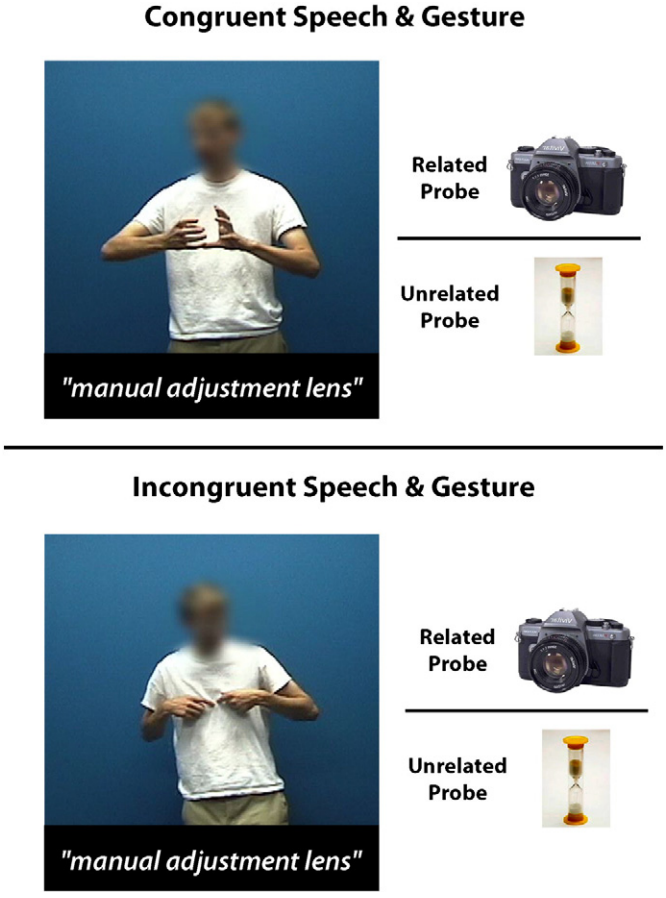

Fig. 1. Discourse primes were comprised of either congruent (top) or incongruent (bottom) speech and gestures, followed by related or unrelated picture probes.

3.3.2. Procedure

Trials began with a title describing the topic of each upcoming discourse segment. Titles were designed to agree with at least the spoken portion of every trial. Since discourse primes were derived from very short snippets of talk extracted from continuous footage, the content of the speaker's utterances could be difficult to grasp in the absences of such preparatory contextual cues. Thus, titles were intended to serve as proxies for the background knowledge that the participant would have accumulated if he or she had been privy to the full conversation.

Following each title, a discourse video was presented at a rate of 30 ms per frame in the center of a computer monitor. After a 50 ms pause, a picture probe appeared in the center of the screen, and remained visible until a response was registered. Between trials, the screen was blank for one second.

Participants were informed that they would be watching a series of short videos in which a man describes various things. They were instructed to read each title silently to themselves, then to watch and listen to each video while holding their index fingers on the A and L keys. When the picture probe appeared, they were told they should press the A or L depending on whether the picture was related or unrelated to the speaker's utterance. Response keys were counterbalanced across individuals. Speed and accuracy were equally emphasized. After

the picture classification task, the Sentence Span and Corsi Block tasks were administered with short breaks in between each test. A debriefing on the purpose of the experiment concluded the session.

3.3.3. Analysis

Response latencies were computed from the onset of the picture probe to the time of the key press. Only correct responses were included in the analysis. Response times exceeding 2.5 standard deviations of the mean for each participant were removed, yielding, on average 2% (SD = 1%) loss of data. RTs by subjects and items and proportions of accurate responses were submitted to repeated measures ANOVA with the factors of discourse congruency (congruent, incongruent) and probe relatedness (related, unrelated). Follow-up contrasts were performed with t-tests.

To assess the relationship between the impact of gestures on discourse comprehension and WM abilities, a multiple regression analysis modeled the speech–gesture congruency effect on response times (incongruent minus congruent, collapsed across relatedness) using span scores from the Corsi Block and Sentence tasks as predictor variables (all measures were standardized by converting to z-scores).

4 . Results

4.1. Accuracy

Analysis revealed that main effects of discourse congruency ($F(1,63) = 8$, $p < 0.05$; congruent more accurate) and probe relatedness ($F(1,63) = 8.5$, $p < 0.05$; unrelated more accurate) were qualified by a two-way interaction ($F(1,63) = 16$, $p < 0.05$). The interaction reflected the presence of a reliable discourse congruency effect for related ($t(63) = 4$, $p < 0.05$), but not unrelated ($t < 1$, n.s.), picture probes. Related picture probes were classified more accurately following discourse primes in which the speech and gestures were congruent than when they were incongruent (see Table 1).

4.2 . Response latencies

Analysis of response latencies revealed main effects of congruency ($F(1,63) = 20.7$, $p < 0.05$; $F_2(1, 139) = 17$, $p < 0.05$; congruent faster) and relatedness ($F(1,63) = 20$, $p < 0.05$; $F_2(1, 139) = 17$, $p < 0.05$; unrelated faster), qualified by an interaction of the two factors ($F(1,63) = 7.8$, $p < 0.05$; $F_2(1, 139) = 6.2$, $p < 0.05$). Follow-up contrasts derived from the analysis by subjects indicated that congruent versus incongruent discourse primes led to more rapid classification of both related ($t(63) = -4.4$, $p < 0.05$) and unrelated ($t(63) = -2.4$, $p < 0.05$) picture probes (Table 1). The interaction occurred because speech–gesture congruency impacted participants' interpretation of the related probes to a much greater degree than the unrelated ones (Fig. 2).

4.3 . Individual differences

Table 2 presents descriptive statistics on Corsi Block Span and Sentence Span, along with the discourse congruency effect obtained by subtracting participants' mean response latencies to picture probes following congruent primes from those following incongruent ones. In keeping with normative studies of the Sentence Span task (Shah & Miyake, 1996), Pearson's correlation tests indicated Sentence Span scores were uncorrelated with our measure of visuo-spatial working memory (Table 2). See Supplementary Fig. A for histograms of the range and distribution of scores on these two WM tasks.

Table 1
Mean accuracy and response latencies on the picture probe classification task.

	Congruent		Incongruent	
	Mean accuracy	Mean RT (ms)	Mean accuracy	Mean RT (ms)
Related	0.95 (0.006 SE)	1184 (60 SE)	0.92 (0.007 SE)	1347 (83 SE)
Unrelated	0.95 (0.005 SE)	1074 (51 SE)	0.96 (0.004 SE)	1131 (56 SE)

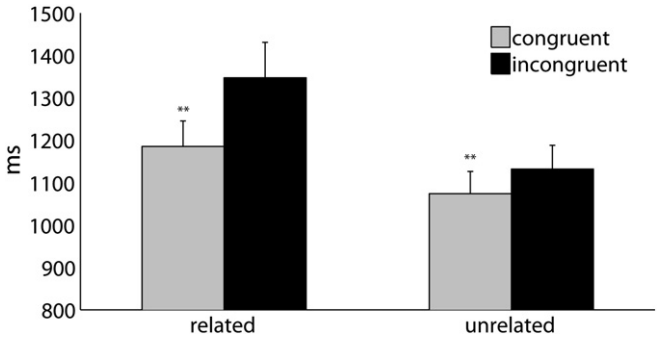

Fig. 2. Mean response latencies (in ms) for picture probe classification in Experiment 1. Error bars reflect 95% confidence intervals. Both related and unrelated picture probes were classified more rapidly when preceded by congruent versus incongruent discourse. However, the magnitude of this discourse congruency effect was much larger for related items.

As noted in the Methods section, the Congruency effect was modeled with multiple regression using both Corsi Block and Sentence Span as predictor variables. Beta weights were .27 for Corsi Block Span and −0.15 for Sentence Span, but only the former was significant ($t = 2.2$, $p < 0.05$). The impact of speech–gesture congruency increased linearly with Corsi Block Span, with larger span scores predicting a larger difference in response times to picture probes primed by congruent versus incongruent speech and gestures (Fig. 3).

5 . Discussion

Experiment 1 was intended to explore the relationship between participants' sensitivity to co-speech iconic gestures and the capacity of their verbal and visuo-spatial WM systems. Results suggest first, that the picture probe classification task was indeed a valid index of participants' sensitivity to iconic gestures, and, second, that visuo-spatial WM helps mediate speech–gesture integration. We briefly discuss each of these points below.

5.1. Picture probe classification task

In keeping with a similarly motivated study by Kelly et al. (2010), results of the picture probe classification task suggest that language comprehenders naturally exploit the information in iconic gestures to help them interpret the accompanying speech. Participants' responses to picture probes were faster and more accurate following the congruent discourse primes than the incongruent ones. Moreover, as predicted, speech–gesture congruency impacted the comprehension of related picture probes to a greater extent than unrelated ones, arguing against the possibility that observed discourse congruency effects were an artifact of video editing. Rather, it suggests that listeners were

Table 2
In Experiment 1, the speech–gesture congruency effect (incongruent minus congruent) on comprehension of picture probes was positively related to Corsi block, but not sentence, span scores.

	Zero-order *r*		
Variable	Congruency effect	1	2
1. Corsi block span	0.26*	–	–
2. Sentence span	−0.14	0.03	–
Mean	162	7.8	3.5
SD	291	1.2	0.9
Min	−118	5.0	1.5
Max	1840	9.0	5.0
Median	82	8.0	3.5

* $p < 0.05$.

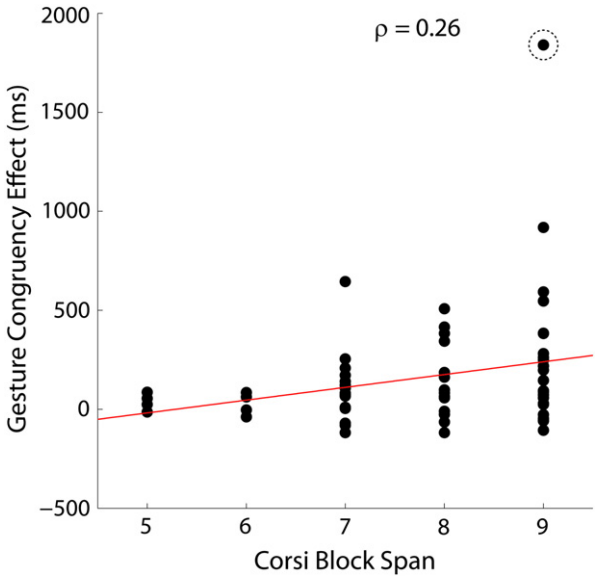

Fig. 3. The effect of speech–gesture congruency on classification times of related pictures (incongruent minus congruent) is positively related to Corsi block span. This correlation remains reliable even when the extreme values associated with the circled data point are removed from the overall sample.

sensitive to differences in the semantic content of the discourse primes. On congruent trials, gestures provided additional information above and beyond what was expressed through speech, perhaps leading to more visuo-spatially rich representations of speaker meaning than was possible on incongruent trials.

Notably, unrelated probes overall were classified more rapidly than related ones. Given that most theories of semantic priming hinge around the basic premise that probe stimuli are processed more easily when related versus unrelated stored knowledge is pre-activated during the priming phase of a trial, this result is counterintuitive. However, it is not unprecedented. A number of studies have reported either negative priming (Alario, Segui, & Ferrand, 2000; Tree & Hirsh, 2003) or no priming (Barrett & Michael, 1990; Holcomb & McPherson, 1994) of responses to related versus unrelated picture probes. Because discourse presented in the current study typically described detailed objects and scenes, it is possible that judging related items required more fine grained analysis (and hence longer decision times) than unrelated ones. For instance, in Fig. 1, correct classification of the unrelated probe simply requires the determination that it does not depict a camera or other optical device, whereas classification of the related probe requires not only that a camera be discerned, but also that the lens of the camera be qualified as one suitable for manual adjustment.

5.2. Individual differences

The magnitude of the speech–gesture congruency effect on discourse comprehension was positively related to Corsi Block span — that is, the impact of speech–gesture congruency on comprehension of picture probes tended to be larger in individuals who were able to remember and reproduce longer block sequences. By contrast, no relationship was detected between sensitivity to speech–gesture congruency and verbal WM abilities. This pattern of outcomes is consistent with the visuo-spatial resources hypothesis, as individuals with high visuo-spatial WM capacity were those who were the most sensitive to co-speech iconic gestures.

One limitation to Experiment 1, however, is its correlational nature. While it is tempting to infer that participants with greater visuo-spatial WM capacity were more sensitive to iconic gestures because visuo-spatial resources are recruited for speech–gesture integration, alternative explanations could be advanced. One possibility is that the observed relationship is simply a coincidence — an artifact of our sample. Statistical analysis suggests this is unlikely, but replication of the findings of Experiment 1 would certainly strengthen the case for the visuo-spatial resources hypothesis. A more serious concern, however, is the possibility that the causally mediating factors for observed speech–gesture congruency effects went unmeasured, except to the extent that they correlate with scores on the Corsi Block task. These concerns about the true nature of the relationship between sensitivity to iconic gestures and visuo-spatial WM capacity motivated the next series of experiments, in which we directly manipulated demands on visuo-spatial and verbal WM resources to examine their impact on sensitivity to information conveyed by iconic gestures.

6 . Experiment 2

In Experiment 2, we further explored possible visuo-spatial contributions to speech–gesture integration through a dual task paradigm designed to tax visuo-spatial WM concurrently with discourse and picture probe processing. The logic of the dual task paradigm is that performance deficits result when two tasks share the same resources (e.g., Wickens, 1980). Accordingly, we hypothesized that a secondary task that draws heavily on visuo-spatial resources will result in diminished capacity to integrate speech and gestures, leading to a smaller congruency effect relative to conditions that do not require intensive visuo-spatial rehearsal.

7. Materials and methods

7.1. Participants

60 new volunteers from the UCSD community (44 female) gave informed consent and received academic course credit for participation. All participants were fluent English speakers.

7.2. Materials, design, and procedure

The primary task was identical to the picture probe classification task used in Experiment 1, requiring participants to view discourse primes and make relatedness judgments to pictures of discourse referents. The secondary task involved remembering a sequence of locations in a two-dimensional grid. Each trial began with the encoding phase of the secondary (Spatial Recall) task in which a fixation cross was displayed for 1 s, followed by a sequence of either one (*low memory load*) or four (*high memory load*) dots distributed pseudo-randomly on a 4×4 grid. After a half second pause, all elements of the primary task (title, discourse prime, and picture probe) were presented according to the same timing parameters used in Experiment 1. However, instead of button presses, participants registered their responses with the mouse (since the secondary task also relied on the mouse as a response device). 500 ms after each relatedness judgment, a blank 4×4 grid appeared, and participants clicked the mouse in the boxes where they remembered seeing target items at the beginning of the trial. Immediately following the final mouse click, written feedback on the Spatial Recall task (either "Correct" or "Incorrect") was shown on the monitor for half a second. Participants were encouraged to respond as quickly and accurately as possible on both the primary and secondary tasks. They were also encouraged to visually rehearse target locations as a memory strategy. After the dual task session, Corsi block and sentence span tests were administered to all participants.

7.3. Analysis

For the primary task, only trials involving correctly classified picture probes were included in the analysis, irrespective of accuracy on the secondary task. Picture probe accuracy and trimmed response latencies (2.5 standard deviations) were analyzed with repeated measures

ANOVA using the factors of speech–gesture congruency and memory load. An average of 2.5% (SD = 0.6%) of the data was lost due to trimming. When warranted by a two-way interaction between these factors, follow-up t-tests were performed. Likewise, a similar repeated measures ANOVA was performed for accuracy scores on the spatial recall test on the secondary task.

As in Experiment 1, the relationship between WM abilities and performance on primary and secondary tasks were evaluated with multiple regression models. On the primary task, the dependent variable was the magnitude of the discourse congruency effect on response latencies collapsed across memory load (incongruent minus congruent). For the secondary task, the dependent variable was participants' accuracy rate on the spatial recall task collapsed across memory load. Both measures were z-normalized and modeled using z-normalized span scores on the Corsi Block and Sentence Span tasks as predictor variables.

8 . Results

8.1. Secondary task accuracy (spatial recall)

As expected, superior recall of target locations was observed in low (93.0%, SD 7.4%) versus high (75%, SD 18%) load trials (memory load main effect: ($F(1,59) = 115$, $p < 0.05$)). Discourse congruity was not significant either as a main effect or in interaction with memory load.

Bivariate correlation coefficients in Table 3 confirm that Corsi Block Span, but not Sentence Span was correlated with overall accuracy on the spatial recall task. The relative import of visuo-spatial versus verbal WM capacity on secondary recall was tested by constructing a multiple regression model in which the dependent measure was overall accuracy on the spatial recall task with Corsi Block and Sentence Span scores as predictors. Beta values in the model suggest the Corsi Block Span ($\beta = 0.36$, $t = 3.0$, $p < 0.01$) was more predictive than Sentence Span ($\beta = 0.14$; $t = 1.1$, n.s.), as the latter was not a significant predictor. Results suggest that the secondary task worked as intended to increase demands on visuo-spatial WM.

8.2. Primary task performance (picture probe classification)

8.2.1. Accuracy

Participants produced more correct judgments of picture probes when discourse primes were comprised of congruent speech and gestures (96%, SD 4%) than when they were incongruent (93%, SD 5.3%) (discourse congruency main effect: $F(1,59) = 15.0$, $p < 0.05$). Accuracy on the picture probe task was also modulated by memory load, as there were somewhat higher accuracy rates in the single target (95.5%, SD 5.4%) than the multiple target (93.8%, SD 6.8%) condition ($F(1,59) = 12.0$, $p < 0.05$). However, there was no evidence of a congruency × load interaction ($F < 1$, n.s.) (Table 4).

Table 3
Descriptive statistics for the following measures in Experiment 2: Gesture congruency effect in ms (subtracting mean response times to congruent trials from incongruent ones, collapsed across levels of WM load), overall accuracy (percent correct) on the secondary spatial recall task, Corsi block span scores, and sentence span scores.

	Zero-order *r*			
Variable	Congruency effect	1	2	3
1. Spatial recall	0.20	–	–	–
2. Corsi block span	0.34*	0.40*	–	–
3. Sentence span	0.18	0.19	0.1	–
Mean	140	84%	7.0	3.7
SD	240	10%	1.3	0.8
Min	−427	55%	5.0	1.5
Max	852	99%	9.0	5.0
Median	148	86%	7.0	3.5

* $p < 0.05$.

8.2.2. Response latencies

As in Experiment 1, congruent discourse primes led to faster classification of pictures than incongruent controls (congruency main effect: $F(1,59) = 20$, $p < 0.05$; $F_2(1,83) = 11.7$, $p < 0.05$). High versus low memory load trials tended to elicit faster responses (load main effect: $F(1,59) = 2.6$, $p = 0.1$; $F_2(1,83) = 7.6$, $p < 0.05$). Further, a congruency × load interaction ($F(1, 59) = 4.6$, $p < 0.05$; $F_2 < 1$, n.s.) indicated that the magnitude of the discourse congruency effect varied as a function of memory load (Table 4). Follow-up t-tests revealed a reliable speech–gesture congruency effect for low memory load trials ($t(59) = 4.6$, $p < 0.05$), along with a non-significant trend toward a congruency effect on high load trials: $t(59) = 1.8$, $p = 0.07$. The interaction presumably reflects the fact that the magnitude of the speech–gesture congruency effect was reduced when participants were required to remember multiple targets (Fig. 4).

8.2.3. Individual differences

Table 3 shows descriptive statistics and bivariate correlation coefficients for Corsi Block Span, Sentence Span, and the Congruency Effect defined as the difference between response latencies to pictures following congruent discourse primes subtracted from those to probes following incongruent primes (collapsed across load), as well as accuracy on the (secondary) Spatial Recall task.

A multiple regression model was used to investigate the relative import of Corsi Block Span and Sentence Span scores for the discourse congruency effect as described above. As in Experiment 1, Corsi Block Span was a significant predictor in the model ($\beta = 0.26$*, $t = 2.1$, $p < 0.05$), but Sentence Span was not ($\beta = 0.19$, $t = 1.5$, n.s.)

9. Discussion

The goal of Experiment 2 was to evaluate speech–gesture integration under the duress of a concurrent secondary task expected to tax visuo-spatial resources (i.e., remembering grid locations). As expected, accuracy of location recall was positively related to performance on a separate test of visuo-spatial, but not verbal, WM capacity. Participants with larger Corsi Block Spans tended to recall more grid locations on both high and low load trials. This finding suggests that the secondary memory task did indeed engage WM resources specific to the visuo-spatial modality, rather than domain general attentional abilities.

Further, participants in Experiment 2 were affected by discourse primes in a similar manner to those in Experiment 1. Congruent versus incongruent speech and gestures led to faster and more accurate assessments of picture probe relatedness. Moreover, the magnitude of this effect increased linearly with Corsi Block Span, but not Sentence Span, scores, in keeping with the idea that individuals with superior visuo-spatial WM exhibit enhanced sensitivity to the semantic properties of depictive gestures.

The novel aspect of Experiment 2 is the finding that memory load can impact sensitivity to gesture meaning. When participants were tasked with remembering four items, the difference in response latencies to related pictures primed by congruent versus incongruent gestures was substantially attenuated relative to the same trials paired with memory sets involving only a single item. Consistent with the visuo-spatial resource hypothesis, these data suggest visuo-spatial WM is engaged during speech–gesture integration. When available resources in this modality were reduced by the secondary task, the congruent gestures exerted less of an effect on participants' understanding of picture probes.

Replicating our findings from Experiment 1, Experiment 2 also revealed that the impact of discourse congruity was related to visuo-spatial WM capacity, as indexed by Corsi Block Span. To illustrate this, Fig. 5 plots the result of regressing the congruency effect in high load trials (i.e. the difference of response latencies to related pictures preceded by congruent discourse primes subtracted from incongruent ones) against Corsi Block Span scores. This analysis reveals that individuals with smaller span scores were more likely to exhibit little or no benefit

Table 4
Mean accuracy and response latencies on the picture probe classification task when visuo-spatial (Experiment 2) and verbal (Experiment 3) WM load were manipulated.

		Experiment 2 Visuo-spatial load		Experiment 3 Verbal load	
		Mean accuracy	Mean RT (ms)	Mean accuracy	Mean RT (ms)
Congruent speech and gesture	Low load	0.97 (0.006 SE)	1622 (49 SE)	0.95 (0.007 SE)	1714 (79 SE)
	High load	0.94 (0.008 SE)	1619 (66 SE)	0.93 (0.010 SE)	1623 (79 SE)
Incongruent speech and gesture	Low load	0.95 (0.008 SE)	1826 (67 SE)	0.92 (0.010 SE)	1884 (91 SE)
	High load	0.92 (0.009 SE)	1696 (61 SE)	0.90 (0.012 SE)	1775 (84 SE)

from congruent gestures on the high load trials. Further, among those individuals whose sensitivity to speech–gesture congruency persisted even under the stress of high memory loads, a near linear increase can be seen in the magnitude of the congruency effect relative to Corsi Block Span scores. As would be expected under the visuo-spatial resources hypothesis, the individuals with smaller span scores were the least sensitive to gestures under the high load.

Of course, further empirical support for the visuo-spatial resources hypothesis could be garnered in a number of ways. For example, it is presently unknown whether the visuo-spatial resources hypothesis would be favored if a more direct, online measure of speech–gesture integration were used, or if a different modality of probe were presented (e.g. lexical items instead of pictures). Moreover, to demonstrate the importance of visuo-spatial resources over other demands exerted by the dual task methodology, Experiment 3 utilized a similar paradigm with comparable demands on executive control processes as Experiment 2, but employed a secondary task designed to tax the phonological loop.

10. Experiment 3

Experiment 3 examines the impact of a secondary verbal WM load on speech–gesture integration. A new cohort of volunteers was presented with a similar paradigm to that employed in Experiment 2. Participants were asked to remember spoken digit sequences consisting of either one or four items during the same picture classification task used in the preceding studies. It is widely believed that this type of recall task engages the phonological loop, as digits are thought to be maintained in immediate memory through verbal rehearsal (reviewed in A. Baddeley, 2003). Given our previous findings implicating visuo-spatial, but not verbal, WM abilities in speech–gesture integration, we predicted that unlike Experiment 2, interpretation of discourse primes would not be modulated by recall load.

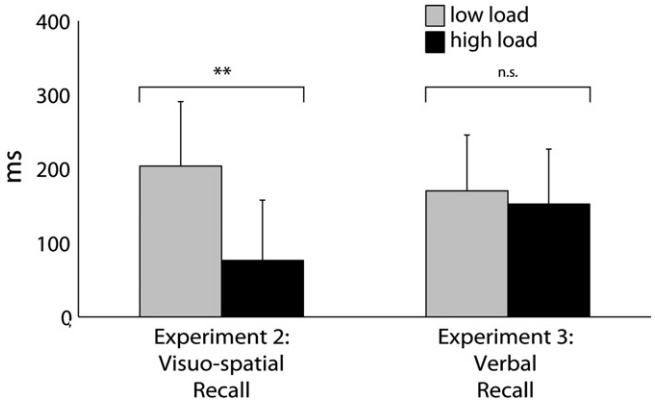

Fig. 4. The mean effect of speech–gesture congruency (incongruent minus congruent) on picture probe classification times when the secondary task involves either a low (single target) or high (4-target) memory load. Under a high secondary visuo-spatial load, the speech–gesture congruency effect is attenuated (Experiment 2). Conversely, under high verbal load, the congruency effect remains of comparable magnitude to that observed on low load trials (Experiment 3).

11. Materials and methods

11.1. Participants

56 new volunteers from the UCSD community (37 female) gave informed consent and received academic course credit for participation. All participants were fluent in English.

11.2. Materials, design, procedure, and analysis

The primary task was identical to that used in Experiment 2. For the secondary task, participants were asked to remember sequences of spoken numbers. At the outset of each trial, a series of digitized audio files containing either one (*low memory load*) or four (*high memory load*) spoken numbers ranging from one to nine was presented while a central fixation cross remained on the computer screen. After completion of the primary picture classification task, an array of randomly ordered digits from 1–9 appeared, and participants clicked the mouse on the numbers that they remembered hearing in the order that they were presented. As before, feedback on secondary task accuracy was provided at the end of each trial, and participants were encouraged to respond as quickly and accurately as possible on both tasks. They were also encouraged to verbally rehearse target numbers. Following the main experiment, Corsi block and sentence span tests were administered.

As in Experiment 2, only responses to correctly classified picture probes were used for analysis, irrespective of accuracy on the secondary task. On average, 1% (sd = 0.1%) of trials was lost due to trimming. The impact of digit load on accuracy and speed of response in the primary task was assessed with a two-way, repeated measures ANOVA using the factors of speech–gesture congruency and memory load.

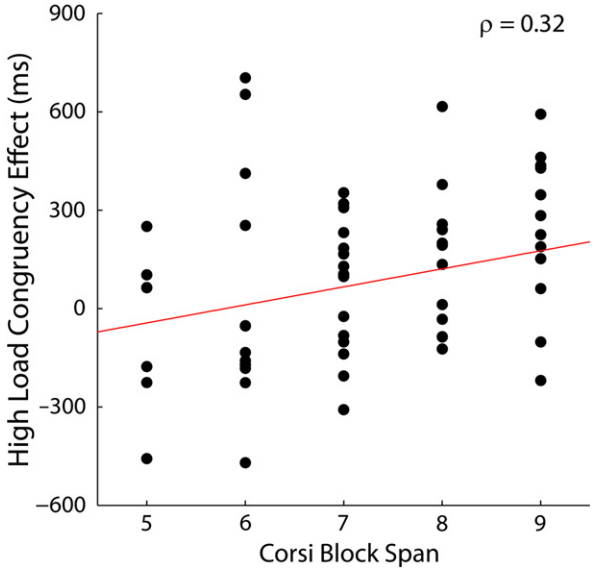

Fig. 5. The effect of speech–gesture congruency on related picture classification times (incongruent minus congruent) on high load trials increases linearly with Corsi block span across individuals.

Additionally, we modeled the relationship between WM abilities and both primary and secondary task performance through multiple regression. The dependent variables were the magnitude of the discourse congruency effect on picture classification times (irrespective of memory load) for the primary task, and rate of accurate digit recall (collapsed across memory load) for the secondary task. As before, span scores on the Corsi Block and Sentence Span tasks served as predictor variables. All measures were normalized.

12. Results

12.1. Secondary task accuracy (verbal recall)

Unsurprisingly, digits were recalled more accurately on low (97%, SD 5%) versus high (89%, SD 10%) load trials (memory load main effect: ($F(1,55) = 54.7$, $p < 0.05$)). No main effect of speech–gesture congruity or interaction with memory load was obtained. Importantly, however, Sentence Span, but not Corsi Block Span, scores were correlated with digit recall accuracy (Table 5). Further, the multiple regression model using Corsi Block and Sentence Span scores as predictor variables revealed that verbal WM capacity (as measured by the Sentence Span task) was much more predictive of digit recall accuracy ($\beta = 0.31$, $t = 2.4$, $p < 0.01$) than visuo-spatial abilities (reflected in Corsi Block Span performance) ($\beta = 0.19$; $t = 1.5$, n.s.), which were not a significant predictor. This pattern of outcomes suggests that the secondary digit recall task did indeed draw on verbal WM resources.

12.2. Primary task performance (picture probe classification)

12.2.1. Accuracy

As in Experiment 2, picture probes were judged more accurately following congruent (94%, SD 6%) than incongruent (91%, SD 8%) primes (discourse congruency main effect: $F(1,55) = 10.3$, $p < 0.05$). Picture probes were also judged more accurately when the secondary memory load contained one item (93%, SD 7%) versus four items (92%, SD 8%) (memory load main effect: $F(1,55) = 6.5$, $p < 0.05$). Again, no congruency × load interaction was found ($F < 1$, n.s.) (Table 4).

12.2.2. Response latencies

As expected, probes were classified faster when preceded by congruent versus incongruent discourse primes (congruency main effect: $F(1,55) = 29$, $p < 0.05$; $F_2(1,77) = 12$, $p < 0.05$). Also, faster responses were found once again on high versus low memory load trials (load main effect: $F(1,55) = 9.4$, $p < 0.05$; $F_2(1,77) = 7$, $p < 0.05$) (Table 4). Crucially, though, no congruency × load interaction was detected (F's < 1, n.s.) (Fig. 4). In the multiple regression model, neither Corsi Block ($\beta = 0.01$; $t = 0.07$, n.s.) nor Sentence Span ($\beta = 0.07$; $t = 0.5$, n.s.) scores were found to be predictive of sensitivity to multi-modal discourse meaning (as measured by the impact of congruent versus incongruent speech and gestures on picture probe classification times).

Table 5
Descriptive statistics for the following measures in Experiment 3: Gesture congruency effect, secondary verbal recall accuracy, Corsi block span scores, and sentence span scores.

	Zero-order *r*			
Variable	Congruency effect	1	2	3
1. Verbal recall	0.04	–	–	–
2. Corsi block span	0.00	0.17	–	–
3. Sentence span	0.07	0.30*	0.08	–
Mean	158	93%	7.2	3.6
SD	221	6%	1.3	0.9
Min	−643	75%	4.0	1.5
Max	672	100%	9.0	5.0
Median	157	95%	7.0	3.5

* $p < 0.05$.

13. Discussion

Experiment 3 examined the relationship between verbal working memory abilities and multi-modal discourse comprehension through a dual task paradigm similar to that employed in Experiment 2. Instead of target locations, participants held number sequences in immediate memory while judging the relatedness of picture probes following segments of discourse containing congruent versus incongruent speech and gestures. As expected, some outcomes of this study paralleled findings from Experiment 2. For instance, just as Corsi Block Span scores were previously found to predict accuracy of spatial recall, the present analysis revealed a positive predictive relationship between performance on the Sentence Span task and digit sequence recall — that is, individuals who could recall larger numbers of sentence final words also reproduced digit sequences more accurately after classifying picture probes. This result is important, as it indicates that the digit load manipulation in Experiment 3 likely engaged verbal WM resources in keeping with experimenters' intentions.

With respect to the primary task, congruent speech and gestures once again led to faster and more accurate classification of related pictures relative to incongruent counterparts, as expected. Also in keeping with Experiment 2, high versus low memory loads resulted in a speed-accuracy trade-off such that picture probes tended to be classified more rapidly, but less accurately on trials involving high loads. This pattern of outcomes may reflect the increased demand placed by large memory loads on a limited capacity executive system.

Crucially, though, Experiment 3 differed from the previous visuo-spatial load paradigm in that the magnitude of the congruency effect was not modulated by verbal load. When participants' verbal WM systems were taxed with large digit sets, their sensitivity to gesture meaning was not adversely impacted, as reflected in the comparable magnitudes of the speech–gesture congruency effect in both the high and low memory load conditions. This finding further bolsters the view that visuo-spatial, but not verbal, WM resources are important for speech–gesture integration. However, the outcomes obtained in Experiments 2 and 3 could also be attributed simply to differences in the overall difficulty of the two types of secondary tasks. In other words, recalling series of target locations may have exacted a substantially greater toll on attentional resources than recalling digit sequences, leading to the reduced speech–gesture congruency effect in the high visuo-spatial versus verbal load conditions. To evaluate this possible confound, a fourth experiment was conducted.

14. Experiment 4

To compare within subjects the attentional demands of the secondary cognitive load manipulations in this study, a new dual task paradigm was created. The primary task involved a conjunctive visual search task analogous to that developed by Treisman and Gelade (1980). This task has been successfully utilized by other behavioral researchers (Hermer-Vazquez & Spelke, 1999) with normative objectives similar to those of the present study. Participants scanned visual displays in search of single target L's that were embedded in a field of distractor T's. On each trial, participants were also expected to remember and recall either a series of four grid locations, in keeping with the parameters of Experiment 2, or four digits in sequence, as in Experiment 3. If these visuo-spatial and verbal recall procedures engage executive resources in a comparable fashion, then we would expect similar response times and accuracy rates in the primary task irrespective of secondary recall modality, even as target detection is challenged by increasing quantities of distractors. Alternatively, one of the two recall modalities could prove measurably more difficult than the other. In this case, it would be crucial to demonstrate that the impact of secondary task modality is stable across increasing levels of primary task difficulty.

15. Methods

15.1. Participants

71 UCSD volunteers (44 female) were awarded course credit for participation in this study. All participants were fluent in English and gave informed consent.

15.2. Materials, design, procedure, and analysis

64 total trials were presented. On half, the secondary task involved the high load version of the visuo-spatial recall task used in Experiment 2, whereas the remainder involved the high load version of the verbal digit recall task from Experiment 3. The primary task involved visual displays containing either seven (small set) or eleven items (large set) arranged in various configurations. On half of the trials, items consisted exclusively of distractors (the letter T), whereas on the remainder, a single target (the letter L) replaced one of the distractors.

Each trial began with a preparatory cross, followed by a sequence of either dots distributed at different locations on a 4 × 4 grid on the computer monitor or recordings of spoken numbers played over the audio system. While keeping either visuo-spatial or verbal items active in immediate memory, participants were asked to scan the displays of distractors and press YES if a target was present or NO if not. Next, either a blank 4 × 4 grid appeared, or a row of numbers. Participants were asked to click with their mouse on the locations or numbers that they remembered in the order of their presentation. Finally, written feedback on recall accuracy (viz. "Correct" or "Incorrect") was shown on the screen.

Analysis was confined only to trials characterized by the presence of a target, as well as accurate performance on both the primary and secondary tasks. Mean response latencies and accuracy were compared between conditions using a 2 × 2 repeated measures ANOVA with the factors of secondary recall modality (visuo-spatial versus verbal) and primary set size (seven versus eleven items). Response times were trimmed within 2.5 standard deviations of the mean, leading to a loss of 2.3% of data.

16. Results

On average, 95% of targets were correctly detected. No main effects of or interactions between secondary recall modality or primary set size were detected (all F's < 1, n.s.). With respect to response times, target detection was reliably slower with eleven (mean: 2251 ms; SD: 645) versus seven distractors (mean: 1994 ms; SD: 441), as expected ($F(1, 70) = 31$, $p < 0.05$). Intriguingly, a main effect of secondary recall modality indicated that verbal recall (mean: 2228 ms; SD: 593) resulted in slower target detection relative to the visuo-spatial task (mean: 2017 ms; SD: 520) ($F(1,70) = 16$, $p < 0.05$). However, as the number of distractors on the primary task increased, the impact of secondary recall modality did not change, as no interaction between secondary recall modality and primary set size was obtained ($F < 1.5$, n.s.) (Fig. 6).

17. Discussion

The purpose of Experiment 4 was to compare within subjects the overall demands placed on executive resources by the two secondary recall tasks used in this study. We reasoned that if attention is required to bind features of targets and distractors in the visual search task (Treisman & Gelade, 1980), then additional attentional loads incurred by the two types of secondary task should impact target detection — both with respect to the error rate and the length of time to make a response. If visuo-spatial and verbal recall tasks engage executive resources in comparable ways, then we would expect to find comparable accuracy rates and response times irrespective of secondary task modality.

Accuracy results confirmed this prediction. Similar rates of error were observed regardless of whether participants were rehearsing visuo-spatial or verbal materials. Mean target detection times suggested a somewhat different picture, however. On average, when taxed with a verbal memory load, participants detected targets over 200 ms more slowly than when maintaining visuo-spatial materials in immediate memory. This finding suggests the verbal recall task may have exacted a greater toll on executive resources than the visuo-spatial one.

Perhaps the most diagnostic measure of processing difficulty in this paradigm, however, is the magnitude of the set size effect — that is, the increase in reaction time for a target embedded in a larger set of targets. When both attentional and perceptual demands are held constant in a visual search task, set size effects are very similar; when task complexity increases, set size effects become greater (Palmer, 1994). In Experiment 4, perceptual demands of the visual search task were identical, with the only differences arising from the attentional demands of the concurrent memory tasks. As can be seen in Fig. 6, the set size effect was very similar for the two secondary recall tasks suggesting they exerted similar demands on executive processing.

When placed in the perspective of the overall study, results of Experiment 4 suggest that differences in the impact of verbal versus visuo-spatial load on sensitivity to speech–gesture congruency in the previous two experiments are not likely due to differences in the overall difficulty of these two tasks. If anything, verbal recall was *more* challenging than visuo-spatial recall — however, under the stress of high memory loads, the secondary visuo-spatial task led to reduced sensitivity to speech–gesture congruency, whereas the secondary verbal task did not. This pattern of results offers further corroboration of the view that visuo-

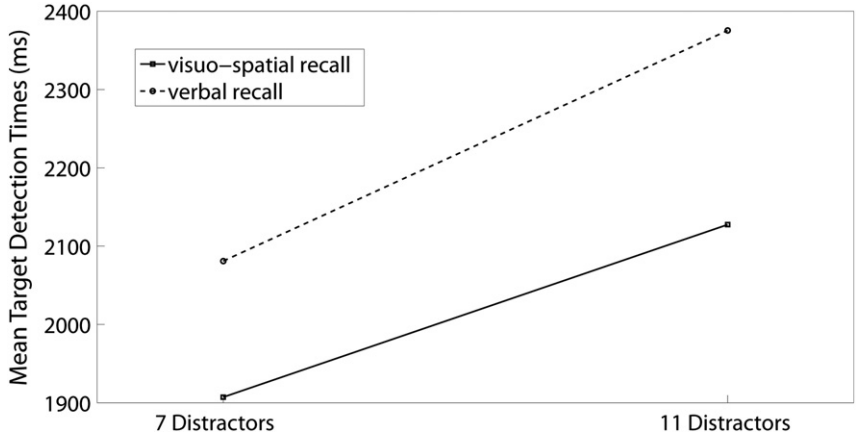

Fig. 6. Mean target detection times in the visual search task (Experiment 4). Overall, targets were detected more rapidly under visuo-spatial versus verbal WM loads. However, the magnitude of the distractor effect was similar for both WM tasks.

spatial resources play a critical role in mediating speech–gesture integration.

18. General discussion

In three experiments, participants classified related pictures more rapidly and accurately when primed by multi-modal discourse with congruent versus incongruent speech and gestures, suggesting first, that people integrate the information conveyed by gestures with that conveyed by the speech, and, second, that our picture probe task offered a reliable index of sensitivity to iconic gestures. Further, Experiments 1 and 2 indicated that the participants who were the most sensitive to the information in iconic co-speech gestures were those with the greatest visuo-spatial WM capacity, consistent with the visuo-spatial resources hypothesis. Moreover, Experiments 2 and 3 demonstrated that increasing demands on participants' visuo-spatial WM system reduced their ability to exploit the information in congruent gestures, whereas comparable loads on verbal WM did not. Again, this pattern of outcomes supports the visuo-spatial resources hypothesis that visuo-spatial WM plays an important role in speech–gesture integration.

Ultimately, though, this conclusion must be drawn with the caveat that this study involved healthy, college-educated adults at an institution that draws from students scoring in the top 15% of their cohort. Although there was little support for the verbal resources hypothesis in our sample, it is possible that the import of verbal WM would be manifested in a more comprehensive study. For instance, among language learners or individuals with clinically impaired language function, low verbal WM abilities might be associated with increased reliance on co-speech gestures.

18.1. Individual differences in gesture production and comprehension

Intriguingly, the findings of the current study are consistent with complementary research in the area of gesture production. Hostetter and Alibali (2007), for instance, have demonstrated high rates of gesturing in individuals with low phonemic fluency scores, but superior spatial visualization skills. Additionally, spatially dominant individuals tended to express information in gesture that was non-redundant with speech more often than those with matched verbal and spatial skills or weak spatial skills (Hostetter & Alibali, 2011). These findings suggest that spatial abilities contribute not only to listeners' understanding of discourse containing gestures, but also to speakers' propensity to produce them. In light of this idea, it is even possible that individuals who gesture more in speaking are also the ones who are impacted the most during comprehension.

On the other hand, a somewhat different pattern of outcomes has also been reported — namely, that visual and spatial WM, as well as mental visualization abilities, were negatively related to rate of gesturing during speech (Chu, Meyer, Foulkes, & Kita, 2014). According to this pattern of results, we would expect a dissociation between individuals who tend to produce gestures while speaking versus those who tend to integrate gestures made by their interlocutors with concurrent speech. Certainly, further research is needed to elucidate the relationship between the factors influencing gesture production and comprehension.

Finally, it should be noted that Wagner, Nussbaum, & Goldin-Meadow (2004) utilized a dual task paradigm in which they examined the impact of a discourse production task with and without accompanying gestures on a concurrent memory task that involved either verbal or visuo-spatial items. In contrast to the present study, which suggests that concurrent visuo-spatial WM load had a detrimental impact on the comprehension of gestures, Wagner et al. (2004) found that gesture production had a beneficial effect on memory for both verbal and visuo-spatial items. Besides the contrast between gesture comprehension and production, one key difference between the present study and that by Wagner et al. (2004) is that discourse in the present study concerned descriptions of concrete objects and actions, whereas participants in Wagner et al.'s study were describing how to factor quadratic equations. Visuo-spatial WM may be more important for understanding gestures about imageable content than abstract topics such as mathematics.

18.2. Visuo-spatial resources hypothesis and the relationship between gesture and space

Observed support for the visuo-spatial resources hypothesis in multi-modal discourse comprehension is in keeping with a broad range of results that point to a close connection between gesture and spatial cognition in both communication and reasoning (see Alibali, 2005 for a review). Producing solution-related gestures during learning, for instance, has been linked to superior retention of math concepts relative to simply stating solution-related goals (Cook, Mitchell, & Goldin-Meadow, 2008). The tendency to gesture during explanations of problem solving strategies has been linked to better performance on a mental rotation task relative to non-gesturers (Ehrlich, Levine, & Goldin-Meadow, 2006), as well as improved subsequent performance in children who were initially unable to solve math puzzles (Broaders, Cook, Mitchell, & Goldin-Meadow, 2007). Beyond impacting task performance, gesturing has also been found to influence problem solving strategies themselves, promoting the use of spatial strategies — even when those strategies are less optimal than available non-spatial alternatives (Alibali, Spencer, Knox, & Kita, 2011).

Gestures have also long been claimed to help people express spatially encoded concepts in a verbal medium (Kita, 2000; McNeill, 1992). Gestures have been shown to help speakers maintain imagistic information in memory (de Ruiter, 1998; Morsella & Krauss, 2004; Wesp, Hess, Keutmann, & Wheaton, 2001). A second benefit for speakers may stem from the reduced communicative load that depictive gestures afford during exchanges about spatial relations. In support of this view, it has been demonstrated that both older children and adults gesture frequently when recounting the locations of hidden objects (Sauter, Uttal, Alman, Goldin-Meadow, & Levine, 2012). Studies in which participants are asked to communicate spatial or imagistic information suggest speakers act as though their gestures have beneficial effects, as gesture rates are higher when addressees are visible than when they are not (Alibali & Don, 2001; Emmorey & Casey, 2002). In fact, effects of addressee visibility are specific to the production of iconic gestures, as comparable effects on the production of rhythmic beat gestures have not been found (Alibali, et al., 2001).

While the bulk of gesture research has centered on their production, connections have also been drawn between spatial cognition and gesture comprehension. Tasks with a strong spatial component, such as reproducing abstract line drawings from a partner's instructions (Graham & Argyle, 1975), or retrieving blocks based on details of their visual characteristics (McNeil, Alibali, & Evans, 2000), have been shown to yield superior performance when participants can both see their interlocutors' gestures and hear accompanying speech, as compared to cases when only speech is presented. Likewise, features such as relative position, speed, and shape were all understood more effectively when a speaker's gestures were visible versus absent during descriptive narratives (Beattie & Shovelton, 2001). Indeed, a recent meta-analysis of 63 studies (Hostetter, 2011) has revealed that relative to abstract content, gestures that depict spatial, and particularly motor, aspects of discourse referents reliably benefit communication. Such findings suggest that iconic gestures positively impact the comprehension of speech that concerns visual, spatial, and action oriented content.

18.3. Gesture and embodied cognition

The discovery that visuo-spatial, but not verbal, abilities mediate the impact of gestures on discourse comprehension is consistent with existing research on gesture processing and speech–gesture integration. 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.

Grounded theories of language have advanced the view that mental simulations of this type are part of everyday language comprehension and reasoning (Barsalou, 2008). 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 versus a mismatched probe depicting an eagle in a nest (Zwaan, Stanfield, & Yaxley, 2002), as would be expected if listeners were mentally simulating imageable aspects of the sentence's meaning. Similarly, when prompted to conceptualize an object from an internal (*driving a car*) or external (*washing a car*) perspective, adults have 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, & Kaschak, 2004). In light of findings such as these, gestures may be viewed as material prompts or scaffolding that can enhance mental simulation processes associated with language comprehension.

18.4. Gesture in everyday settings

An important finding of the current study is that not everyone was affected by gestures in the same way, as individuals with greater visuo-spatial working memory capacity were more likely to take advantage of the information conveyed by the gestures. Results of the present study thus have potential implications for education. Previous research has demonstrated that instruction involving co-speech gestures as opposed to speech alone can lead to improved learning in areas such as navigation (Glenberg & Robertson, 1999), quantitative reasoning (Church, Ayman-Nolley, & Mahootian, 2004), math problem solving (Singer & Goldin-Meadow, 2005), geometry (Valenzeno, Alibali, & Klatzky, 2003), and Japanese[1] (Kelly, McDevitt, & Esch, 2009). Current findings suggest that to achieve optimal multi-modal communication, it is necessary to take individual differences into account. In keeping with Gardner's theory of multiple intelligences (Gardner, 1993), it may be useful in educational settings for learners to experience new concepts presented through a variety of approaches that include varying discourse styles.

Results of the picture classification task are in keeping with prior work suggesting that participants are sensitive to the semantic properties of iconic gestures and their relationship to the accompanying speech (Wu & Coulson, 2005, 2007). Unlike other paradigms assessing speech–gesture integration (Kelly, Ozyurek, & Maris, 2010), the picture task did not explicitly instruct participants to attend to the speaker's gestures or to the semantic congruency between his gestures and the content of his utterances. In fact, it was possible to perform this task while ignoring the gestures altogether. Moreover, unlike much research on gesture comprehension that involves materials generated by actors, the stimuli used in the present study were derived from videos of spontaneous discourse — and consequently involved both dysfluent speech characteristic of conversational settings, along with gestures that were less "crisp" than those in constructed materials. Results of the present study thus reinforce claims in the literature that language users spontaneously exploit the information in co-speech gestures (Kelly et al., 2010), and suggests such claims may extend to speech–gesture integration processes that operate in everyday discourse settings.

One aspect of the materials here, however, was quite unnatural – namely, the blurring of the speaker's face – and may have induced more attention on his gestures than one would find in a real world exchange. We find this unlikely, however, in view of the fact that participants were not explicitly instructed to attend to gestures, and could have performed the task without recourse to the information that they expressed. Parametric variation of the intelligibility of speech and blurring of the face during speech suggests the visibility of the face has a large impact on language processing when the intelligibility of speech is low, but is far less consequential for the sort of readily intelligible speech used in the present study (Scott et al., 2002). Further, a number of studies suggest that speech–gesture integration is a fairly obligatory process (Kelly, Creigh, & Bartolotti, 2009; Kelly et al., 2010).

Finally, the design of the present study does not allow the possibly facilitative effects of congruent speech and gestures to be distinguished from potentially inhibitory effects when these two modalities are incongruent. However, measures of observed facilitation versus inhibition are inevitably contingent on properties of the baseline that one selects, and may be vulnerable to over- or underestimation depending on the dimensions along which baseline stimuli differ from experimental ones (Jonides & Mack, 1984). Moreover, the primary goal of the present work was not to add to extant research demonstrating either the benefit (Holle & Gunter, 2007) or cost (Bernardis, Salillas, & Caramelli, 2008; Kelly et al., 2004) exerted by co-speech gestures. Because the present study explored how sensitivity to gesture meaning relates to other cognitive abilities, it was deemed preferable to examine the more powerful comparison of congruent versus incongruent gestures.

19. Conclusion

In conclusion, the present study demonstrates an important role for visuo-spatial resources in multi-modal discourse comprehension. In three experiments, healthy adults classified related picture probes more rapidly when primed by discourse with congruent versus incongruent gestures. The novel finding advanced here is that not all listeners are impacted equally by gestures. In particular, these data suggest visuo-spatial WM capacity plays a more important role in mediating speech–gesture integration than verbal abilities. Consistent with a causal role for visuo-spatial WM in speech–gesture integration, the data further suggest that increasing demands on visuo-spatial WM has a detrimental impact on the ability to exploit the information in iconic gestures. These findings are in keeping with the idea that co-speech iconic gestures promote image-based simulations of the meaning of an utterance, and may help explain how gestures aid comprehension.

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.actpsy.2014.09.002.

Acknowledgments

This work was supported by a grant to SC from the NSF (#BCS-0843946). Special thanks go to Rebecca Dai, Jordan Davison, and Marguerite McQuire for their contributions.

References

Alario, F. X., Segui, J., & Ferrand, L. (2000). Semantic and associative priming in picture naming. *The Quarterly Journal of Experimental Psychology*, *53A*(3), 741–764.
Alibali, M. W. (2005). Gesture in spatial cognition: Expressing, communicating, and thinking about spatial information. *Spatial Cognition and Computation*, *5*(4), 307–331.
Alibali, M. W., & Don, L. S. (2001). Children's gestures are meant to be seen. *Gesture*, *1*(2), 113–127.
Alibali, M. W., Heath, D. C., & Myers, H. J. (2001). Effects of visibility between speaker and listener on gesture production: Some gestures are meant to be seen. *Journal of Memory and Language*, *44*(2), 169–188.
Alibali, M. W., Spencer, R. C., Knox, L., & Kita, S. (2011). Spontaneous gestures influence strategy choices in problem solving. *Psychological Science*, *22*(9), 1138–1144.

[1] Interestingly, learning word meanings in a foreign language is aided by hand gestures that accompany instruction, but learning to perceive phonetic contrasts in a second language is not (Hirata & Kelly, 2010).

Baddeley, A.D. (2003). Working memory: Looking back and looking forward. *Nature Reviews. Neuroscience*, *4*, 829–839.

Baddeley, A.D., & Hitch, G. J. (1974). Working memory. In G. A. Bower (Ed.), *Recent advances in learning and motivation, vol. 8.* (pp. 47–90). New York: Acadamic Press.

Barrett, S. E., & Michael, D. R. (1990). Event-related potentials and the semantic matching of pictures. *Brain and Cognition*, *14*, 201–212.

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

Beattie, G., & Shovelton, H. (2001). An experimental investigation of the role of different types of iconic gesture in communication: A semantic feature approach. *Gesture*, *1*(2), 129–149.

Berch, D. B. (1998). The Corsi block-tapping task: Methodological and theoretical considerations. *Brain and Cognition*, *38*, 317–338.

Bernardis, P., Salillas, E., & Caramelli, N. (2008). Behavioural and neurophysiological evidence of semantic interaction between iconic gestures and words. *Cognitive Neuropsychology*, *25*(7–8), 1114–1128.

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

Broaders, S.C., Cook, S. W., Mitchell, Z., & Goldin-Meadow, S. (2007). Making children gesture brings out implicit knowledge and leads to learning. *Journal of Experimental Psychology: General*, *136*(4), 539–550.

Chu, M., Meyer, A., Foulkes, L., & Kita, S. (2014). Individual differences in frequency and saliency of speech-accompanying gestures: The role of cognitive abilities and empathy. *Journal of Experimental Psychology: General*, *143*(2), 694–709.

Church, R. B., Ayman-Nolley, S., & Mahootian, S. (2004). The role of gesture in bilingual education: Does gesture enhance learning? *International Journal of Bilingual Education and Bilingualism*, *7*(4), 303–319.

Conway, A.R., Kane, M. J., Bunting, M. F., Hambrick, D. Z., Wilhelm, O., & Engle, R. W. (2005). Working memory span tasks: A methodological review and user's guide. *Psychonomic Bulletin & Review*, *12*(5), 769–786.

Cook, S. W., Mitchell, Z., & Goldin-Meadow, S. (2008). Gesturing makes learning last. *Cognition*, *106*(2), 1047–1058.

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

de Ruiter, J. (1998). *Gesture and speech production.* Netherlands: Catholic University of Nijmegen.

Ehrlich, S. B., Levine, S.C., & Goldin-Meadow, S. (2006). The importance of gesture in children's spatial reasoning. *Developmental Psychology*, *42*(6), 1259–1268.

Emmorey, K., & Casey, S. (2002). Gesture, thought, and spatial language. *In Spatial Language* (pp. 87–101). Springer Netherlands.

Gardner, H. (1993). *Multiple intelligencies: The theory in practice.* New York: Basic Books.

Glenberg, A.M., & Robertson, D. A. (1999). Indexical understanding of instructions. *Discourse Processes*, *29*(1), 1–26.

Graham, J. A., & Argyle, M. (1975). A cross-cultural study of the communication of extra-verbal meaning by gestures. *International Journal of Psychology*, *10*(57–67).

Hadar, U., & Krauss, R. K. (1999). Iconic gestures: The grammatical categories of lexical affiliates. *Journal of Neurolinguistics*, *12*, 1.

Hadar, U., & Pinchas-Zamir, L. (2004). The semantic specificity of gesture: Implications for gesture classification and function. *Journal of Language and Social Psychology*, *23*(2), 204–214.

Hermer-Vazquez, L., & Spelke, E. S. (1999). Sources of flexibility in human cognition: Dual-task studies of space and language. *Cognitive Psychology*, *39*, 3–36.

Hirata, Y., & Kelly, S. D. (2010). Effects of lips and hands on auditory learning of second-language speech sounds. *Journal of Speech, Language, and Hearing Research*, *53*, 298–310.

Holcomb, P., & McPherson, W. B. (1994). Event-related brain potentials reflect semantic priming in an object decision task. *Brain and Cognition*, *24*, 259–276.

Holle, H., & Gunter, T. C. (2007). The role of iconic gestures in speech disambiguation: ERP evidence. *Journal of Cognitive Neuroscience*, *19*(7), 1175–1192.

Hostetter, A. B. (2011). When do gestures communicate? A meta-analysis. *Psychological Bulletin*, *137*(2), 297–315.

Hostetter, A. B., & Alibali, M. W. (2007). Raise your hand if you're spatial: Relations between verbal and spatial skills and gesture production. *Gesture*, *7*(1), 73–95.

Hostetter, A. B., & Alibali, M. W. (2011). Cognitive skills and gesture–speech redundancy: Formulation difficulty or communicative strategy. *Gesture*, *11*(1), 40–60.

Hostetter, A. B., & Hopkins, W. D. (2002). The effect of thought structure on the production of lexical movements. *Brain and Language*, *82*(1), 22–29.

Jonides, J., & Mack, R. (1984). On the cost and benefit of cost and benefit. *Psychological Bulletin*, *96*(1), 29–44.

Kelly, S. D., Creigh, P., & Bartolotti, J. (2009). Integrating speech and iconic gestures in a stroop-like task: Evidence for automatic processing. *Journal of Cognitive Neuroscience*, *22*(4), 683–694.

Kelly, S. D., Kravitz, C., & Hopkins, M. (2004). Neural correlates of bimodal speech and gesture comprehension. *Brain and Language*, *2004*(89), 253–260.

Kelly, S. D., McDevitt, T., & Esch, M. (2009). Brief training with co-speech gesture lends a hand to word learning in a foreign language. *Language and Cognitive Processes*, *24*(2), 313–334.

Kelly, S. D., Ozyurek, A., & Maris, E. (2010). Two sides of the same coin: Speech and gesture mutually interact to enhance comprehension. *Psychological Science*, *21*(2), 260–267.

Kita, S. (2000). *How representational gestures help speaking.* Cambridge: Cambridge UP.

Krauss, R. M., Dushay, R. A., Chen, Y., & Rauscher, F. (1995). The communicative value of conversational hand gestures. *Journal of Experimental Social Psychology*, *31*, 533–552.

Lavergne, J., & Kimura, D. (1987). Hand movement asymmetry during speech: No effect of speaking topic. *Neuropsychologia*, *25*(4), 689–693.

McNeil, N., Alibali, M. W., & Evans, J. L. (2000). The role of gesture in children's comprehension of spoken language: Now they need it, now they don't. *Journal of Nonverbal Behavior*, *24*(2).

McNeill, D. (1992). *Hand and mind.* Chicago: Chicago University Press.

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

Morsella, E., & Krauss, R. M. (2004). The role of gestures in spatial working memory and speech. *American Journal of Psychology*, *117*(3), 411–424.

Ozyurek, A., Willems, R. M., Kita, S., & Hagoort, P. (2007). On-line integration of semantic information from speech and gesture: Insights from event-related brain potentials. *Journal of Cognitive Neuroscience*, *19*(4), 605–616.

Palmer, J. (1994). Set-size effects in visual search: The effect of attention is independent of the stimulus for simple tasks. *Vision Research*, *34*(13), 1703–1721.

Sauter, M., Uttal, D. H., Alman, A. S., Goldin-Meadow, S., & Levine, S.C. (2012). Learning what children know about space fom looking at their hands: The added value of gesture in spatial communication. *Journal of Experimental Child Psychology*, *111*(4), 587–606.

Scott, S. K., Rosen, S., Spitsyna, G., Faulkner, A., Neville, L., & Wise, R. J. S. (2002). *The neural basis of cross-modal enhancement in speech perception: A PET study.* (Abstract No. 354. 17).

Shah, P., & Miyake, A. (1996). The separability of working memory resources for spatial thinking and language processing: An individual differences approach. *Journal of Experimental Psychology: General*, *125*, 4–27.

Singer, M.A., & Goldin-Meadow, S. (2005). Children learn when their teacher's gestures and speech differ. *Psychological Science*, *16*(2), 85–89.

St. George, M., Mannes, S., & Hoffmant, J. E. (1997). Individual differences in inference generation: An ERP analysis. *Journal of Cognitive Neuroscience*, *9*(6), 776–787.

Straube, B., Green, A., Weis, S., & Kircher, T. (2012). A supramodal neural network for speech and gesture semantics: An fMRI study. *PloS One*, *7*(11), http://dx.doi.org/10.1371/journal.pone.0051207.

Tree, J. J., & Hirsh, K. W. (2003). Sometimes faster, sometimes slower: associative and competitor priming in picture naming with young and elderly participants. *Journal of Neurolinguistics*, *16*, 489–514.

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

Valenzeno, L., Alibali, M. W., & Klatzky, R. (2003). Teacher's gestures facilitate students' learning: A lesson in symmetry. *Contemporary Educational Psychology*, *29*, 187–204.

Wagner, S. M., Nusbaum, H., & Goldin-Meadow, S. (2004). Probing the mental representation of gesture: Is handwaving spatial? *Journal of Memory and Language*, *50*(4), 395–407.

Waters, G. S., & Caplan, D. (1996). The measurement of verbal working memory capacity and its relation to reading comprehension. *The Quarterly Journal of Experimental Psychology. A*, *49*(1), 51–79.

Wesp, R., Hess, J., Keutmann, D., & Wheaton, K. (2001). Gestures maintain spatial imagery. *The American Journal of Psychology*, *114*, 591–600.

Wickens, C. D. (1980). The structure of attentional resources. *Attention and Performance VIII*, *8*.

Willems, R. M., Ozyurek, A., & Hagoort, P. (2007). When language meets action: The neural integration of gesture and speech. *Cerebral Cortex*, *17*, 2322–2333.

Wu, Y. C., & Coulson, S. (2005). Meaningful gestures: Electrophysiological indices of iconic gesture comprehension. *Psychophysiology*, *42*(6), 654–667.

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.

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