The role of speech-gesture congruency and delay in remembering action events

Alexia Galati1 and Arthur G. Samuel1,2,3

1Department of Psychology, Stony Brook University, Stony Brook, NY, USA
2Basque Center on Cognition Brain and Language, Donostia-San Sebastian, Spain
3IKERBASQUE, Basque Foundation for Science, Bilbao, Spain

When watching others describe events, does information from their speech and gestures affect our memory representations for the gist and surface form of the described events? Does our reliance on these memory representations change over time? Forty participants watched videos of stories narrated by an actor. Each story included three target events that differed in their speech-gesture congruency for particular actions (congruent speech/gesture, incongruent speech/gesture, or speech with no gesture). Participants had to reproduce target event sentences, prompted after delays of 2, 6, or 18 minutes. Seeing gestures, either congruent or incongruent, led to better gist recall (more mentions of the target action, more gestures for the target action, and more complete target events) compared to not seeing gestures. However, seeing incongruent gestures sometimes led to reproductions of the incongruent gestures, particularly after short delays, as well as inaccuracies in speech. Our results suggest that over time people increasingly rely on multimodal gist-based representations and rely less on representations that include surface and source information about speech and gesture.

Keywords: Speech-gesture processing; Gist memory; Gesture.

Correspondence should be addressed to Alexia Galati, PO BOX 59628, 4011 Limassol, Cyprus. E-mail: alexia.galati@gmail.com

This material is based upon work supported by NSF under grants IIS-0527585 and ITR-0325188, and by NIH grants R01-51663 and R01-059787. We would like to thank Richard Gerrig, Susan Brennan, and our colleagues from the Gesture Focus Group for helpful discussion. We are grateful to Drew Boudreau for serving as our actor, Matthew Belevich, Gennadiy Ryklin, and Marina Khan for help with coding, and Donna Kat for her enormous technical assistance. We also thank two anonymous reviewers for suggestions that significantly improved this paper.

© 2010 Psychology Press, an imprint of the Taylor & Francis Group, an Informa business
http://www.psypress.com/lcp DOI: 10.1080/01690965.2010.494846
Memory for the surface form of discourse is short-lived, except under special circumstances (e.g., Johnson-Laird & Stevenson, 1970). The semantic content (or gist) of discourse material is better retained over time than the actual words and syntactic form (or the verbatim wording) for sentences (Brewer & Nakamura, 1984; Kintch, Welsch, Schmalhofer, & Zimny, 1990; Sachs, 1967) and conversations (Hjelmquist & Gidlund, 1985). The representations constructed during discourse processing are generally thought to retain the gist or meaning of sentences but not their surface form, giving rise to gist-based “mental models” (Johnson-Laird, 1983) or “situation models” (Kintsch, 1998).

Such gist representations can be elaborated by more than one modality since information that is experienced multimodally is learned and remembered better than if it is experienced in only one modality (for a review see Shams & Seitz, 2008). For example, pictures can improve the memory for words (Anderson & Bower, 1973; Paivio, 1986) and the learning of text (Carney & Levin, 2002). Stimuli with multisensory pasts are also more accurately discriminated: previously viewed images are differentiated better from new images when their prior presentation was in an auditory–visual pair than only visual (Murray et al., 2004). Moreover, the visual articulatory movements of the face can help speech comprehension (Schwartz, Berthommier, & Savariaux, 2004) and memory for sentences (Thompson, 1995).

Speech-accompanying gestures, like other kinds of visual information, have the potential to facilitate the processing of speech and elaborate mental representations. In this paper, we examine how gestures can contribute to surface and gist-based memory representations. We first establish that people can routinely extract information from gestures, since this is a prerequisite for gestures to affect gist-based memory representations. We then review studies on the relationship between memory and gesture to motivate our main research question: whether over time source information about gesture becomes less available, while semantic information that is extracted from gesture and is incorporated in a gist-based memory representations fades less rapidly.

**PEOPLE ROUTINELY EXTRACT INFORMATION FROM GESTURES**

Some studies have approached the question of whether people process information from gestures by asking whether people can extract information from gestures in isolation. These studies generally demonstrate that the informative value of gestures is low relative to that of speech (Hadar & Pinchas-Zamir, 2004; Krauss, Morrel-Samuels, & Colasante, 1991). While in these studies people extract more information from speech than from
gestures presented on their own, they can nonetheless identify the meaning of gestures better than chance (see also Kendon, 1994 for a related critique). For example, when seeing gestures without sound, participants were about 76% accurate at choosing between two possible meanings (Krauss et al., 1991, Experiment 1) and they generated interpretations for gestures that others successfully classified in a forced-choice task, especially for semantic categories like actions and locations (Krauss et al., 1991, Experiment 2).

Similarly, people can choose the word that best reflects the gesture’s meaning more often than chance, even when seeing the gesture on its own without any speech context (Hadar & Pinchas-Zamir, 2004). However, finding that people can extract information from gestures when seeing them in isolation does not directly address how they extract information in the context in which gestures are normally processed: with speech.

Other studies have addressed the information value of gestures more directly by comparing how well people comprehend or remember statements with and without gestures. In some studies, when participants watched videos of descriptions with gesture along with speech, they were not significantly better at choosing the referent of the description from two alternatives compared to when they heard the descriptions alone, suggesting that the contribution of gestures is small (Krauss, Dushay, Chen, & Rauscher, 1995). But assessing the semantic components of participants’ responses, instead of using a forced-choice task, reveals that they do extract substantially more information when seeing gestures, especially about semantic categories like the relative position, speed, and shape of entities (Beattie & Shovelton, 1999, 2001).

People are in fact more likely to rely on an interlocutor’s gestures under certain circumstances. For example, people benefit more from seeing others’ gestures when the material being described is difficult to encode. When watching speakers describe the shape of two-dimensional objects of high and low codability, participants’ drawings of the objects were more accurate when the descriptions were accompanied by gestures than when they were not, and the effect was greater for objects of low codability (Graham & Argyle, 1975). People also use gestures when speech is pragmatically ambiguous. For example, they are more likely to interpret a pragmatically ambiguous utterance as an indirect request rather than a declarative statement, when it is accompanied by a pointing gesture than when it is not (Kelly, Barr, Church, & Lynch, 1999).

People are more likely to rely on gestures not only when faced with phonetic and pragmatic ambiguity, but also when the amount of signal interference increases. In a study in which participants chose the meaning of descriptions presented at four signal-to-noise ratios, the presence of gestures significantly increased comprehension scores as noise was introduced (Rogers, 1978). Similarly, representational gestures (gestures
depicting semantic content through the hands’ movement, shape, or placement) presented with speech in noise increased the retention of critical items compared to vague gestures or no gestures (Riseborough, 1981, Experiment 3). As the noise level was increased, performance dropped in all conditions except when participants saw representational gestures. This advantage of representational over nonrepresentational gestures was also demonstrated in another study: sentences accompanied by representational gestures were more likely to be recalled correctly and recognised accurately than sentences accompanied by nonrepresentational gestures or without gestures (Feyereisen, 2006, Experiments 1a & 1b).

Also, children are more likely to rely on gestures with increased ambiguity of speech. They rely on gestural cues, like pointing gestures towards objects, when speech is phonetically ambiguous (Thompson & Massaro, 1994) or pragmatically ambiguous, as when using deictic words like here, there, this, and that (Tfouni & Klatzky, 1983). When seeing pointing gestures, children are more likely to identify the correct referent and are less likely to commit deictic errors (i.e., to choose a different token of the same type than the one indicated). They can also extract information from combinations of nonredundant information in speech and gesture conveying conceptual task-related information (Kelly & Church, 1997). Some studies suggest that there may be a developmental change in the ability to incorporate information from gesture with information from speech, with age facilitating our ability to extract information from gesture. Often, increased ability to extract information from gesture leads to positive outcomes: for example, in an ambiguous phonetic context, 9-year old children are more likely to rely on gestures than 4-year olds (Thompson & Massaro, 1994). But occasionally, increased ability to extract information from gestures can lead to less desirable outcomes: for example, adults are worse than children at recalling speech that was accompanied by mismatching gestures (Kelly & Church, 1998).

More recently, ERP studies have provided evidence that people extract information from gestures and rapidly integrate it with information from speech. These studies have demonstrated that gestures impact on-line brain measures that are associated with semantic processing. Many of the studies addressing the integration of speech and gesture use a mismatch paradigm, where incongruent information is presented in the two modalities (Holle & Gunter, 2007; Kelly, Kravitz, & Hopkins, 2004; Özyürek, Willems, Kita, & Hagoort, 2007; Wu & Coulson, 2005). These ERP studies usually assess the N400 component in the EEG signal, showing greater difficulty (a larger N400) when integrating an incongruent gesture than a congruent gesture with the speech context. For example, Holle and Gunter (2007) found that when ambiguous homonyms (e.g., ball) that were accompanied by a representational gesture supporting one of their meanings were followed
by a target word (e.g., game or dance), the target word was associated with a larger N400 when it was incongruent with the earlier gesture–homonym combination than when it was congruent.

Overall, there is growing support for the claim that people extract information from gestures and can incorporate it with information from speech during speech comprehension. However, what remains an open question is to what extent people incorporate information extracted from gesture into their memory representations and how long lasting these resulting memory representations are.

SPEECH AND GESTURE AS TWO SOURCES OF MEMORY

Evidence for extracting information from gesture during speech processing, reviewed in the previous section, leads to our view that memory representations have inputs from both speech and gesture and are multimodal in nature. We hypothesise that to the extent that people incorporate information from speech with information from gesture, they will over time increasingly rely on a multimodal gist-based representation and rely less on representations that include surface and source information about the two modalities.

This hypothesis is consistent with current frameworks in memory research, such as fuzzy-trace theory (Brainerd & Reyna, 2002; Brainerd, Reyna, & Brandse, 1995; Reyna & Brainerd, 1995; Reyna & Kiernan, 1994), which posit that distinct gist and verbatim traces are formed in parallel when material is encoded. In such a framework, correct recollection is supported by both of these traces, and forgetting involves their gradual fragmentation, which is higher for verbatim than for gist traces. Therefore, in this view, people are more likely to access verbatim information immediately after encoding and rely on gist representations after a delay. In the current study, in order to examine whether over time people increasingly rely on multimodal gist representations, we manipulate to-be-remembered material in terms of the congruency of semantic features of speech and gesture.

Evidence from a recognition task suggests that the concurrent processing of speech and gesture leads to poor retention of the surface form of speech, supporting the view that speech and gesture form multimodal gist-based representations (Cutica & Bucciarelli, 2008, Experiments 2 & 3). Participants who saw an actor narrating either with or without (congruent and redundant) gestures later judged whether test sentences were part of the story; the test sentences could be verbatim, paraphrases, or new sentences. Participants were more likely to misidentify paraphrases as the studied statements when the statements had been accompanied by gestures than when they had not. Similarly, people don’t retain the surface form of gestures
either, insofar as they do not seem to identify gestures they have seen before. In forced-choice recognition tests on whether video clips were new or old, participants who had initially only heard the audio of the clips were just as accurate in distinguishing between new and old clips as those who watched video clips with audio, and more accurate than participants who watched the video clips without sound (Krauss et al., 1991, Experiments 3 & 4). While this was taken as evidence that the contribution of gestures to communication is small, an alternative interpretation is that information from gesture becomes integrated with speech to form a single multimodal representation, leading to poorer recognition of the surface form of both speech and gestures.

In light of evidence that surface information from speech alone becomes inaccessible more quickly than gist information (Brewer & Nakamura, 1984; Hjelmquist & Gidlund, 1985; Kintch et al., 1990; Sachs, 1967) and that processing speech with gestures leads to elaborated gist-based representations (Cutica & Buccarelli, 2008), gestures may potentially bolster gist information, in particular, against the passage of time.

To date this has not been examined directly. However, one study that was focused primarily on gist memory did not find evidence for gestures bolstering gist information against memory loss. Church, Garber, and Rogalski (2007) examined whether statements are remembered better over time if they had been accompanied by gestures. Participants watched video stimuli of statements with or without gestures that were nonredundant with speech and wrote their recollections in response to written prompts, either immediately after watching the videos or 30 minutes later. Statements that were accompanied by gestures were remembered better than those without gesture and both types of statements were remembered more poorly over time. Critically, there was no interaction between the presence of gesture and delay: watching statements with gesture did not reduce memory loss over time compared to watching statements without gesture. Using somewhat different procedures, we test whether memory for both gist and verbatim information is modulated by not only the presence of gesture, but also its semantic congruency to speech.

Manipulating speech-gesture congruency can afford insight into underlying memory representations (Cassell, McNeill, & McCullough, 1999; McNeill, Cassell, & McCullough, 1994). Cassell and colleagues found that speakers’ retellings of a story more often omitted events that, in the original stimulus, were accompanied by incongruent gestures (conveying semantic content that was incompatible with that of speech) relative to events with congruent gestures (conveying the same semantic content as speech), or complementary gestures (conveying compatible but nonredundant content with respect to speech). Speakers’ retellings had more departures from speech for events with incongruent gestures than congruent gestures, but
there were departures from speech for complementary gestures as well. Thus, incongruent information from gesture affected both the memory representations’ strength, as indicated by participants’ omissions of story elements, and their accuracy, as indicated by participants’ confabulations. To determine to what extent the semantic relationship of speech and gesture affects memory, we examine how well people remember events with congruent and incongruent gestures,¹ as Cassell and colleagues (1999) did, and also how well they remember events that were not accompanied by any gestures.

We consider gist memory traces to be abstract representations of semantic content that do not incorporate details of surface form, while verbatim memory traces to include surface form and source information (Reyna & Kiernan, 1994). We focus here on those memory representations that arise after initial encoding of an event (beyond the time span of working memory), but before consolidation through sleep. These types of memories support performance in a wide range of everyday situations. To take just a few examples, this variety of memory would be at work in a classroom when a student invokes a concept that the teacher explained earlier in the lecture, in a business meeting when the speaker refers to an idea that a colleague had introduced at an earlier point, and in a social interaction when speakers revisit events of an anecdote that their conversational partners had narrated at the beginning of the conversation. In these situations, a memory representation for an event or concept may be accessed multiple times after initial encoding but before consolidation. If, for instance, a specific event of a previously narrated anecdote is reinvoked only once by a speaker, then we expect its recall to be subjected to memory loss over time. At the same time, if the anecdote as a whole is accessed repeatedly by the speaker during the conversation, its recall may be subject to hypermnesia, an improvement in recall with each repeated attempt (for a review see Payne, 1987).

¹ Although a potential criticism is that incongruities between verbal and nonverbal information are unnatural or implausible in communication, there is evidence that they are not. Speech-gesture mismatches are prevalent in transitional points of cognitive development (e.g., Church & Golden-Meadow, 1986; Goldin-Meadow, Nusbaum, Garber, & Church, 1993; Perry, Church, & Goldin-Meadow, 1988), and adults use such mismatching gestures to assess children’s knowledge (Alibali, Flevares, & Goldin-Meadow, 1997; Goldin-Meadow & Singer, 2003; Goldin-Meadow, Wein, & Chang, 1992). Mismatches can also occur between people’s speech and facial displays, and observers readily interpret them. For example, positively valued utterances paired with negatively valued facial expressions and vocal qualities were judged by respondents to be sarcastic, while negatively valued utterances paired with positively valued nonverbal displays were judged to involve joking (Bugental, Kaswan, & Love, 1970). As Bavelas and Chovil (2000) have pointed out, addressees interpret inconsistent pairings of verbal and nonverbal information felicitously assuming that they are intended by the speaker to be part of a single unified message, consistent with Grice’s (1967/1989) cooperative principle. Thus, it is not the case that our participants should be unable or unwilling to interpret speech-gesture incongruities felicitously.
In our study, by manipulating the congruency between speech and gesture and the time between encoding and recall, and by examining both gist and verbatim recall, we gain further insight into how reliance on gist and surface representations arising from processing speech with gestures might change over time. We predict that recall accuracy for target events will differ across congruency conditions: insofar as gestures bolster gist representations there should be a benefit of seeing gestures and insofar as people integrate spatiomotoric features from speech and gesture, there should be a benefit of seeing congruent over incongruent gestures. We expect that both gist and verbatim recall for the target events will worsen over time. However, for gist memories, which involve the semantic processing of speech and gesture, delay should interact with congruency, such that memory for events that were accompanied by congruent gestures should be more stable over time. Finally, we expect that memory loss for the surface form of speech and gesture will be higher than for gist, with participants relying increasingly on multimodal gist representations for the target events (Brainerd & Reyna, 2002; Reyna & Brainerd, 1995).

Our theoretical perspective also makes predictions about the distribution of gestures that participants themselves produce during recall. Participants should gesture more if they have observed a speech-accompanying gesture than if they have not, because spatiomotoric features about the event to be described should be accessible in the multimodal representations. Also, because surface information about the observed gestures should become less accessible over time, with people increasingly relying on multimodal gist traces, we predict that upon observing an event with an incongruent gesture, participants should produce more incongruent gestures at shorter delays than at longer delays; at longer delays, the mismatch of gesture with speech should impair accurate speech report, as people may incorporate inconsistencies in their multimodal gist representations.

Our manipulation of the congruency of speech and gesture becomes especially important here, since producing a congruent gesture for a target event that was accompanied by a congruent gesture does not distinguish whether the observed gesture helped gist memory for that event, verbatim memory, or both. The congruent features of the participant’s gesture could have been generated by either an abstract representation of the semantic content of the event or by surface features of the target event, including the observed congruent gestures. On the other hand, reproducing an incongruent gesture (without a co-occurring inaccuracy in speech) for a target event that was accompanied by an incongruent gesture would reflect verbatim memory for the target event, because both the source (namely, the gestural modality) and surface features of the narrator’s telling are maintained. On the other hand, inaccuracies in speech for these events would suggest that features of the observed incongruent gesture have been incorporated into a gist.
representation. By manipulating congruency along with delay, and analysing participants’ gestures as well as their speech, we gain insight into how surface representations arising from the processing of gesture with speech may attenuate over time while gist-based representations remain relatively stable (particularly, when information from the two modalities is congruent).

METHOD

Participants

Forty-seven students from Stony Brook University whose native language was English participated for research credit. The data of three participants were excluded because of technological failures, and the data of four participants were excluded because they did not follow instructions to look at the computer screen. The data of the remaining 40 participants were analysed.

Materials

Nine target stories and four filler (see below) stories were narrated by an actor. Each story was approximately 2 minutes long. The video of each story was segmented into 15 clips, such that each sentence of the story was a separate video clip.

Each story included three target motion events, always occurring at sentences 4, 9, and 14 of the story. One of the three target events included a target motion verb for which speech and gesture were congruent, conveying the same semantic information. Another target event included a target motion verb for which speech and gesture were incongruent, conveying incompatible semantic information. The remaining target event included a target motion verb that was not accompanied by gesture; the semantic information for the target verb was conveyed only in speech. Appendix 1 shows an example of such a story.

Each target event was prompted for recall either after a short delay (immediately after the target story; approximately 2 minutes), after an intermediate delay (after the story following the target story; approximately 6 minutes), or after a long delay (four stories after the target story; approximately 18 minutes).

Congruent, incongruent, and no gesture events were distributed to the different positions in the story (sentences 4, 9, and 14) and to their delay condition (short, intermediate, and long) through Latin squares. Of the 27 target motion verbs, nine were transitive action verbs (e.g., plug), nine were intransitive manner verbs (e.g., tiptoe), six were intransitive path verbs (e.g., zigzag), and three were a combination of a generic motion verb with an
adverb (e.g., *moves steadily*). All verbs were selected to have a potential mismatch (e.g., *plug–unplug, tiptoe–stomp, zigzag–spiral, move steadily–move shakily*) whether they were accompanied by an incongruent gesture or not.

Stories were presented in one of two orders, the sequence of the nine target stories being reversed in one of the orders. For both orders, the four filler stories followed the target stories to provide the intermediate and long delays for the last four target stories.

**Apparatus**

Participants were tested in a sound-shielded booth. They advanced through video clips and recall trials by pressing a button; presentation of the stimuli was controlled by a Pentium PC outside the booth. The video was presented on a computer display inside the booth and the audio was presented binaurally over high-quality stereo headphones.

**Procedure**

Participants were told that they would be watching videos of stories narrated by an actor and that they had to remember the stories as completely and accurately as possible. They were informed that at different points in the experiment they would be prompted to reproduce sentences from the story. They were told that for each prompt they had to reproduce the single following sentence. They were specifically instructed to reproduce sentences in the exact same words as much as possible; if they could not, they could paraphrase but they should aim for verbatim reproduction whenever possible. When recalling sentences, participants were instructed to treat the experimenter as their audience and were informed that they would not be receiving any feedback regarding the accuracy of their responses. The experimenter always smiled and nodded at participants’ responses. Participants were videotaped during the cued recall sessions.

For each story, participants first saw the title of the story presented for 3 seconds; then, the first video clip of the story was automatically presented. The presentation of the videos was then self-paced, with participants pressing a key to move to the next sentence. Participants watched each story twice, in both passes clicking a key after each video clip. The second presentation of each story was designed to bring recall performance into a reasonable accuracy range.

After watching a story twice, participants engaged in cued recall. For each recall, prompts included the title of the story and video clips of the two sentences prior to the target sentence. Prompts were interspersed throughout the experiment to provide the desired delays. For example, after the fifth story, participants received a *short* delay prompt to recall a target event from the fifth story, an *intermediate* delay prompt to recall a target event from the
fourth story, and a long delay prompt to recall a target event from the first story. Note that for a given story, the prompts after a short, intermediate, and long delay were always different and nonoverlapping, prompting a different target sentence. This procedure avoids repetition of any probed sentence, maintaining the desired manipulation of probe delay. At a more global level, activating the story representation itself could potentially generate hypermnesia.

At the end of the experiment, participants were asked if they had noticed anything out of the ordinary in the stimuli, in order to assess how salient speech-gesture incongruities may have been. Participants were then debriefed.

Speech coding

The first author transcribed the participants’ responses to the 27 target items and coded for whether the target action for the event was mentioned, the extent to which the target event was reproduced completely, and the extent to which the target event was reproduced verbatim. The first two measures were taken to assess gist memory, while the last was taken to assess surface memory.

1. **Mention of target action:** This measure accounted for cases where the target action was realised in a verb phrase in a correctly recalled event; the measure combined correct reproductions and substitutions of the target verb. It involved a judgement of whether the target verb had been omitted, or was otherwise reproduced or substituted (e.g., “makes a boomerang shape” instead of “boomerangs”). Changing the tense or aspect of the target verb, or inserting it in a frame such as “starts to . . .” or “decides to . . .”, did not affect the coding. Since mentioning the target action involved both reproductions and substitutions of the target verb, and considering that participants were instructed to recall the speech of target events verbatim, we took this measure to reflect primarily abstract representations of semantic content (gist memory), though reproductions of the target verb could also reflect verbatim memory.

2. **Complete reproduction of target event:** This measure was intended to capture the degree of gist recall and involved a judgement of whether participants had reproduced all, most, or part of the propositional content of the sentence. This measure taps into gist memory not only for the target action, but also for additional propositional content relevant to the target action. It allows us to assess whether speech-gesture congruency for a particular unit of propositional content (the target action) can affect the gist memory for the target event as a whole.
For a Complete: All judgment, all the content words of the target event’s description had to be reproduced either verbatim or with content words conveying an equivalent amount of detail. Content words included verbs, nouns, adjectives, and adverbs. A Complete: Most judgment was made when 50% or more of the content words of the sentence had been reproduced, or were replaced by content words of equivalent detail, with the remaining content words either omitted or replaced by less detailed content words. For example, if a participant reproduced all the content words of the target event’s description but replaced the target verb with a less informative verb (e.g., goes instead of zigzags), a Complete: Most judgment was made. For a Complete: Part judgment, less that 50% of the content words had to be reproduced or replaced with content words conveying an equivalent amount of detail, with the remaining content words either omitted or replaced by less detailed content words.

3. Verbatim reproduction of target event: This measure attempted to capture verbatim recall and involved a judgement of whether sentences were reproduced fully verbatim, mostly verbatim, partly verbatim, or not verbatim. For a Verbatim: All judgment 100% of the words of the target sentence had to be reproduced. However, substituting nouns with pronouns, using different but appropriate prepositions (e.g., saying “under the impact” instead of “from the impact”), and minor additions (e.g., saying “in order to avoid the rock”) did not affect this judgment. For a Verbatim: Most judgment, over 50% of the words of the target sentence needed to be reproduced. The reproduced words had to be consecutive in one or, at most, two strings. In other words, there could only be one departure from the surface form of the target sentence, which could be either at the beginning or end of the sentence (resulting in one consecutive string of words), or in the middle (resulting in two strings). For a Verbatim: Part judgment, 25 to 50% of consecutive words of the target sentence had to be reproduced. If participants reproduced less than 25% of consecutive words from the target sentence, we judged that the surface form of the sentence had not been reproduced verbatim and assigned the Verbatim: Not judgment.

Reliability

To assess reliability for these measures, we had a second coder, an undergraduate research assistant (blind to the congruency and delay conditions of the target events) redundantly code 25% of the corpus (10 participants). The two coders agreed 95% of the time on whether the target action was mentioned. For the complete reproduction of the target event, we calculated the proportion of times the coders agreed that the participants had reproduced most or all of the propositional content, since
our analyses in the “Results” Section are based on this combined category. Agreement for this measure of reproduction of the target event was 91%. For verbatim reproduction of the target event, we calculated the proportion of times the coders agreed that the participants had reproduced most or all of the words of the target sentence (since, again, our analyses are based on this combined category): agreement was 85%.

**Gesture coding**

The first author also coded for gestures in the subset of correctly recalled events.

First, gestures were identified by considering all hand movements produced by participants and excluding those that were irrelevant hand movements (e.g., scratching one’s nose or adjusting glasses).

Then, gestures were classified as either being representational or not. Representational gestures were defined as gestures depicting semantic content by virtue of handshape, placement, and movement (e.g., Alibali, Heath, & Myers, 2001; also referred to as *iconics* by McNeill, 1992). The process of gesture classification was guided both by the semantic features that were encoded in gesture and by our elicitation protocol, which afforded coders insight into what participants intended to convey. Since we did not have predictions for how congruency or delay would affect nonrepresentational gestures, we did not consider them further.

Of the representational gestures produced, only those that represented the target action were considered since we did not have predictions for how congruency or delay would affect gestures representing other aspects of the target event.

Finally, representational gestures for the target action were classified as being congruent or incongruent with the target verb. If the elements of motion encoded in the gesture (e.g., its manner, path, and direction) were incompatible with those implied by the target verb, the gesture was judged as incongruent. Otherwise, the gesture was judged as being congruent with the target verb. This criterion allowed us to detect potential inaccuracies observed in gesture for events from any of the three congruency conditions.

**Reliability**

To assess reliability for identifying representational gestures for the target action, we had a second coder, an undergraduate research assistant (blind to the congruency and delay conditions of the target events) redundantly code 25% of the corpus (10 participants). The coders agreed 98% of the time on whether the participant produced a gesture for target event, 89% of the time on whether there was a representational gesture for the target action, and
96% of the time on whether a representational gesture for the target action was congruent or not with the target verb.

Analyses

Analyses were conducted of the proportion of correct events with the target action mentioned, correct events with most or all of the propositional content complete, correct events realised mostly or fully verbatim, and correct events including a representational gesture for the target action. Each analysis was a $3 \times 3 \times 2$ ANOVA with speech-gesture congruency (congruent, incongruent, and no gesture) and delay (short, intermediate, and long) as the within-participants factor and with the order of the stories as the between-participants factor. These analyses were conducted with both participants ($F_1$) and target items ($F_2$) as the random variables. There were no reliable effects involving story order and thus results for story order are not reported.

RESULTS

Correct target action

Overall, participants remembered the correct target event quite well, about 83% of the time ($SD = 0.37$). Presenting each story twice during the experiment was successful in getting participants to remember events, which had been one of our initial concerns. Given this solid overall level of recall, we can meaningfully examine recall for the correct target action. This measure should provide the most direct assessment of any effects of congruency and delay on participants’ gist memory. We looked at how often participants realised the correct target action in order to assess whether over time gist memory, as reflected by mentions of the target action, was affected.

Specifically, we predicted that recall accuracy for target actions would differ across congruency conditions, such that there would be a benefit of seeing gestures, insofar as gestures bolster gist representations, and there should also be a benefit of seeing congruent over incongruent gestures, insofar as spatiomotoric features from gesture are incorporated with speech. As shown in Figure 1a, participants were 12% more likely to mention the target action, either by reproducing the target verb or replacing it with a synonym, when seeing a congruent gesture than an incongruent gesture, and 23% more likely to do so when seeing an incongruent gesture than no gesture at all. Indeed, speech-gesture congruency significantly affected how likely participants were to mention the target action, $F_1(2, 70) = 56.78$, $MSE = 0.0061$, $p < .001$; $F_2(2, 18) = 7.53$, $MSE = 0.0076$, $p < .01$. As we predicted, actions from events without gesture were mentioned significantly less frequently than those from events with congruent gestures, $F_1(1, 35) = 113.42$,
Those with incongruent gestures, $F_1(1, 35) = 32.01$, $MSE = 0.0052$, $p < .001$; $F_2(1, 18) = 11.09$, $MSE = 0.0076$, $p < .01$. Thus, seeing a description of an action accompanied by a gesture supported gist memory for the target action, and moreover seeing a congruent gesture helped more than seeing an incongruent gesture, $F_1(1, 35) = 23.80$, $MSE = 0.0030$, $p < .001$; $F_2(1, 18) = 2.22$, $MSE = 0.0076$, $p = .15$.

Second, we expected that speech-gesture congruency would interact with delay, such that seeing a gesture, particularly a congruent one, protected against memory loss over time. While performance overall was relatively flat across delay, $F_1(2, 70) = 0.45$, $MSE = 0.0048$, $ns$; $F_2(2, 18) = 0.11$, $MSE = 0.0076$, $ns$; see Figure 1b, for target actions accompanied by incongruent gestures, those prompted after a short delay were remembered better than those prompted after an intermediate, $F_1(1, 37) = 5.35$, $MSE = 0.17$, $p < .05$; $F_2(1, 6) = 2.18$, $MSE = 0.0076$, $p = .19$ or a long delay, $F_1(1, 37) = 7.70$, $MSE = 0.0095$, $p < .01$; $F_2(1, 6) = 4.66$, $MSE = 0.0076$, $p = .07$, as shown in Figure 1c. There was no such effect of delay for events with congruent or no gestures; notably, recall of the target actions accompanied by congruent gestures was at ceiling. The effect of delay for actions

Figure 1. Mean proportion of target action verbs across congruency (a), across delay (b), and across both factors (c).
accompanied by incongruent gestures seems to be responsible for the interaction between congruency and delay, which was significant only by participants, $F_1(4, 140) = 2.88$, $MSE = 0.0056$, $p < .05$; $F_2(4, 18) = 0.31$, $MSE = 0.0076$, ns. Thus, in line with our predictions, the recall of target actions accompanied by congruent gestures was protected against the effect of delay, remaining at ceiling, while recall of target actions accompanied by incongruent gestures did decline at intermediate and long delays. Somewhat surprisingly, regardless of the delay, not seeing a gesture for the target action at all led to the worst recall for the target action. This suggests that even the incongruent gestures may have initially strengthened the gist representations, as might be the case if people treat gestures as a signal of increased importance of a given portion of the message.

**Complete reproduction of target events**

We also investigated when participants would reproduce most or all of the propositional content of the target event. We considered this measure to reflect gist memory for the target event as a whole. As a function of delay, we examined whether the type of gesture accompanying the target action would affect not only recall of the target action, but also recall for the entire event in which the target action was embedded.

To the extent that gestures for the target action affect gist representations for the target event and for the target action in the same way, we expected that seeing a gesture for the target action would improve how completely events were recalled. Also, to the extent that spatiomotoric features from gesture are incorporated with speech during processing, we expected congruent gestures to have an advantage. Indeed, seeing a gesture accompanying the target action affected not only recall for the target action, as shown in the previous section, but also recall for information of the target event as a whole. As shown in Figure 2a, participants were 10% more likely to reproduce events relatively completely when events were accompanied by a congruent gesture than when they were accompanied by an incongruent gesture, and 14% more likely when they were accompanied by an incongruent gesture than when they were not accompanied by a gesture. Speech-gesture congruency affected how completely participants reproduced the propositional content of target events, $F_1(2, 70) = 27.77$, $MSE = 0.0076$, $p < .001$; $F_2(2, 18) = 2.21$, $MSE = 0.11$, $p = .14$. Participants were more likely to reproduce most or all of the propositional content when they had seen congruent gestures as opposed to no gestures, $F_1(1, 35) = 71.86$, $MSE = 0.0039$, $p < .001$; $F_2(1, 18) = 0.69$, $MSE = 0.11$, ns. Seeing incongruent gestures was also associated with somewhat more complete recall than not seeing any gestures, $F_1(1, 35) = 13.50$, $MSE = 0.0065$, $p < .01$; $F_2(1, 18) = 0.250$, $MSE = 0.11$, ns.
We expected that to the extent that gestures for the target action affect gist representations for the target action and for the target event in the same way, gestures, particularly congruent ones, would reduce any performance decrement over time. However, this was not the case. Unlike our findings for recall of the target action, the interaction of congruency and delay here was not significant, $F_1(4, 140) = 0.23, \text{MSE} = 0.0091, \text{ns}; F_2(4, 18) = 0.10, \text{MSE} = 0.11, \text{ns}$. As Figure 2c shows, in all three congruency conditions recall of the target event’s propositional content was actually worse in the most immediate test than after longer delays, $F_1(2, 70) = 4.43, \text{MSE} = 0.0093, p < .05; F_2(2, 18) = 0.38, \text{MSE} = 0.11, \text{ns}$. Figure 2b illustrates this pattern, collapsing across the congruency manipulation. Participants reproduced the propositional content mostly or fully completely 10% more often after an intermediate delay than after a short delay, and 4% more often after a long delay than after an intermediate delay. After an intermediate or long delay participants were more likely to reproduce content fully than after a short delay [short vs. intermediate: $F_1(1, 35) = 7.04, \text{MSE} = 0.0093, p < .01; F_2(1, 18) = 0.11, \text{MSE} = 0.11, \text{ns};$ short vs. long: $F_1(1, 35) = 5.05, \text{MSE} = 0.0063, p < .05; F_2(1, 18) = 0.02, \text{MSE} = 0.11, \text{ns}$.]

This unexpected pattern probably is due to hypermnnesia. Our design obliged participants to recall information from each story three times during the experiment. Although the probed events were carefully separated in the passages, when participants tried to recall events after an intermediate or
long delay, prior probing could have activated passage information that led to more complete reproduction compared to events prompted after a short delay. An indication that participants accessed target events earlier in the experiment is that in 20 instances in our corpus, target events prompted after an intermediate or long delay (out of 720 in our dataset) had been mentioned during earlier prompting while recalling another target event of the same story, despite our instructions to recall a single event for each prompt. The distinct patterns for recall of target events, vs. target actions, may be instructive: the latter are relatively specific, and less likely to have been activated by prior probes, whereas the former are basic components of the narrative and thus more likely to have been activated when the passage was probed previously (potentially supporting hypermnesia).

**Verbatim reproduction of target events**

We also investigated when participants’ reproduction of target events would be at least mostly verbatim. Recall that, operationally, we considered responses to be mostly or fully verbatim if 50% or more of the words were reproduced verbatim in one or two consecutive strings.

Based on previous research on gist and verbatim memory, we predicted that participants’ performance in recalling events verbatim would be poorer than their gist memory. We have just seen (Figure 2) that when participants recalled the correct event, overall, they reproduced most or all of the propositional content 63% (SD = 0.48) of the time. In contrast, they reproduced it mostly or fully verbatim only 14% (SD = 0.34) of the time. This pattern, as noted, is consistent with studies showing that people are worse at remembering the surface form of passages relative to their gist.

We expected that like gist memory, verbatim memory should be improved by seeing gestures, to the extent that gestures highlight the accompanying speech. Indeed, participants reproduced events mostly or fully verbatim 18% (SD = 0.38) of the time when the events were accompanied by a congruent gesture, 15% (SD = 0.36) when accompanied by an incongruent gesture, and 8% (SD = 0.28) when they were not accompanied by a gesture, $F_1(2, 70) = 5.85$, $MSE = 0.0040$, $p < .01$; $F_2(2, 18) = 1.06$, $MSE = 0.0044$, $ns$. Compared to events with no gesture, participants were more likely to reproduce target events mostly or fully verbatim when they were accompanied by gestures [congruent vs. no gesture: $F_1(1, 35) = 10.91$, $MSE = 0.0027$, $p < .05$; $F_2(1, 18) = 2.91$, $MSE = 0.0044$, $p = .11$; incongruent vs. no gesture $F_1(1, 35) = 6.40$, $MSE = 0.0025$, $p < .05$; $F_2(1, 18) = 2.46$, $MSE = 0.0044$, $p = .13$]. Consistent with our earlier finding that seeing a gesture helped participants reproduce the target event more completely, seeing a gesture also helped them reproduce the event verbatim—gestures elaborated both gist and surface traces.
Finally, although we had predicted that seeing gestures, particularly congruent ones, would prevent memory loss of gist memory (as we found to be the case for recalling the target action), we did not expect such an interaction for verbatim recall: while the semantic processing of congruent gestures may result in more enduring gist representations against time, it need not result in more enduring surface representations. Indeed, none of the interactions between congruency and delay suggested that gestures protected verbatim memory against memory loss. In fact, surprisingly, delay did not affect how likely participants were to reproduce the entire event verbatim: after a short delay participants reproduced the target events mostly or fully verbatim 11% (SD = 0.31) of the time, after an intermediate delay 17% (SD = 0.37), and after a long delay 13% (SD = 0.34); these differences were not significant [short vs. intermediate: \( F_1(1, 35) = 1.70, \text{MSE} = 0.0344, \text{ns} \); \( F_2(1, 18) = 0.62, \text{MSE} = 0.0044, \text{ns} \); intermediate vs. long: \( F_1(1, 35) = 0.37, \text{MSE} = 0.026, \text{ns} \); \( F_2(1, 18) = 0.69, \text{MSE} = 0.0044, \text{ns} \)].

**Gestures encoding target actions**

Since participants could have encoded information about the target event not only in speech but also in their gestures, we examined the proportion of correctly realised target events that included a representational gesture for the target action. Based on the view that observing a gesture for the target action would make spatiomotoric features about the action available (presumably both in gist and surface representations), we predicted that participants would be more likely to produce a gesture for the target action when the target action had been described with a speech-accompanying gesture than when it had not. And insofar as spatiomotoric features from gesture are incorporated with speech in a gist representation, seeing a congruent gesture for the target action should lead to the most elaborated gist representation and increase the likelihood of producing a gesture for the target action. As Figure 3a shows, when participants had seen a congruent gesture they produced a representational gesture for the target action 11% more often than when they had seen an incongruent gesture, and when they had seen an incongruent gesture they produced a representational gesture for the target action 8% more often than when they had not seen a gesture. This difference in the likelihood of producing a representational gesture for the target action across congruency conditions was reliable, \( F_1(2, 70) = 18.70, \text{MSE} = 0.0055, p < .001; F_2(2, 18) = 3.40, \text{MSE} = 0.0042, p = .06 \). As expected, participants were more likely to produce a representational gesture for the target action after seeing a congruent gesture than after not having seen any gesture, \( F_1(1, 35) = 28.90, \text{MSE} = 0.0046, p < .001; F_2(1, 18) = 10.88, \text{MSE} = 0.0042, p < .01 \). Less predictably, they were also more likely to produce a representational gesture after seeing an incongruent gesture...
than after not having seen any gesture, $F(1, 35) = 7.69$, $MSE = 0.0017$, $p < .01$; $F(1, 18) = 2.34$, $MSE = 0.0042$, $p = .14$.

To the extent that spatiomotoric features about the target action would be more vivid and available at shorter delays than at longer delays, we expected that the effect of seeing a gesture would be stronger for shorter delays than longer delays. We did indeed find that, for events prompted after a short delay, seeing a gesture significantly increased the likelihood of producing a representational gesture [no gesture vs. congruent: $F(1, 36) = 22.54$, $MSE = 0.12$, $p < .001$; $F(1, 6) = 29.14$, $MSE = 0.0042$, $p < .01$; no gesture vs. incongruent: $F(1, 36) = 15.73$, $MSE = 0.12$, $p < .001$, $F(1, 6) = 19.51$, $MSE = 0.0042$, $p < .01$]. In terms of an overall effect of delay on gesture production (Figure 3b), the proportion of representational gestures for the target action decreased numerically over time: participants produced a representational gesture for the target action 3% more frequently after a short delay than after an intermediate delay, and 5% more frequently after an intermediate delay than after a long delay. While the overall effect of delay on gesture production was marginal, $F(2, 70) = 2.67$, $MSE = 0.0046$, $p = .08$; $F(2, 18) = 0.56$, $MSE = 0.0042$, $ns$, participants were more likely to produce gestures for the target action after a short than a long delay, $F(1, 35) = 4.88,$
\[ MSE = 0.0032, \ p < .05; \ F_2(1, \ 18) = 1.695, \ MSE = 0.0042, \ p = .21, \] particularly for events with incongruent gestures, \( F_1(1, \ 37) = 6.20, \ MSE = 0.12, \ p < .02, \ F_2(1, \ 6) = 4.75, \ MSE = 0.0042, \ p = .07. \] The difference between intermediate and long delays overall was not significant, though it was for events with no gesture, \( F_1(1, \ 35) = 10.21, \ MSE = 0.0091, \ p < .01, \ F_2(1, \ 6) = 8.92, \ MSE = 0.0042, \ p < .05. \) To summarise, at a short delay seeing a gesture led to more representational gestures for the target action than not seeing a gesture. After long delays, gesture production did not decline when having seen a congruent gesture, but it did when having seen an incongruent one.

**Distribution of incongruent gestures produced by participants**

As we have pointed out, producing a representational gesture for the target action does not tell us whether observing a gesture improved gist memory for the target action or surface memory for the actor producing the gesture. We therefore analysed whether participants produced gestures that were congruent or incongruent relative to the target action verb because this allows us to assess the compatibility between those gestures participants produced and those they observed.

Based on the view that people extract semantic information from gestures, we expected that they would produce more incongruent gestures when they had seen an incongruent gesture than in the other congruency conditions. Moreover, based on the view that people rely on gist representations increasingly over time, we expected that surface information from incongruent gestures would be less accessible at longer delays, with most incongruent gestures being produced at short delays. As Table 1 shows, this was indeed the case: after seeing incongruent gestures, participants produced a sizeable number of incongruent gestures, primarily after a short delay. Still, after seeing incongruent gestures participants in general produced more congruent gestures than incongruent gestures. This was also the case for events without a gesture; for these events, only one incongruent gesture was produced and this was after a long delay. Notably, after seeing congruent gestures, participants never produced incongruent gestures. There was a difference in the distribution of congruent gestures across conditions of congruency and delay: \( \chi^2(4) = 16.4, \ p < .01. \) The distribution of incongruent gestures was not significantly different across conditions of congruency and delay, although its probability was relatively small (\( p = .20, \) Fisher’s exact test).

\[ ^2 \text{From this assessment of a total of 209 representational gestures for the target action, we excluded five quoted incongruent gestures. In these gestures participants reproduced consciously, and not unwittingly, incongruities performed by the actor in the stimulus. Their quotation was marked by a combination of facial displays, laughing, and exaggerated gesturing.} \]
Inaccuracies in the speech and gesture modalities

Finally, we examined the number of departures from the actor’s speech occurring not only in the participants’ gestures, but also in their speech. We expected that, while incongruent gestures would be most likely at short delays upon seeing an incongruent gesture (see Table 1), inaccuracies in speech arising from these incongruent gestures would be more likely at longer delays. If participants rely increasingly on gist representations, then misleading information that was originally presented in gesture and that has presumably been incorporated in gist representations would be more likely to be accessed from these representations at longer delays and reported in speech, despite instructions to reproduce the actor’s speech verbatim. This is consistent with our earlier finding that gesturing the target action is more likely at shorter delays upon having seen a gesture, and that after longer delays gesturing becomes less likely especially upon having seen an incongruent gesture (Figure 3c).

As Table 2 shows, indeed, seeing an incongruent gesture led to more inaccuracies in speech after longer delays. This was also true after participants saw events that were not supplemented by a gesture; again these inaccuracies increased after longer delays. This is consistent with fuzzy-trace theory’s account that gist retrieval can lead to the recall of false details that are consistent with the overall meaning of the event (Brainerd & Reyna, 2002; Reyna & Brainerd, 1995), such as one participant saying “buttons her jacket” instead of “zips up her jacket” when there had been no gesture.
accompanying this target action. Of the incongruent gestures reported in Table 1, most occurred only in gesture upon seeing an incongruent gesture at a short delay, with only three incongruent gestures being accompanied by inaccuracies in speech. Regardless of the delay, events with congruent gestures never led to inaccuracies in speech, just as they never led to incongruent gestures, as we reported in the previous section. That is, overall, inaccuracies occurred only for events with incongruent gestures or with no gestures: upon seeing incongruent gestures, people produced more inaccurate gestures after short delays and made more errors in speech after long delays; upon not seeing any gestures, people’s verbal recall was more inaccurate over time.

Inaccuracies in speech were often in the form of incompatible substitutions of the target verb. Table 3 lists some examples of compatible substitutions (to provide some context for comparison) and all \( n = 12 \) incompatible substitutions of the target verb occurring in our corpus. The remaining four inaccuracies in speech were not in the form of verb substitutions; they involved confabulations that were not specific to the target verb. One such example comes from a participant recalling an event about a character who was moving cautiously due to having an arthritic knee. The event was accompanied by an incongruent gesture, with the actor saying “Michael moves cautiously to avoid any further injury” while moving his hands carelessly from side to side during “moves cautiously”. In recalling this event, the participant said “he feels pain or something and he goes like ‘oooooooh’”, while reproducing the actor’s gesture. In other words, the

| Table 2 |
|---|
| Number of inaccuracies occurring in gesture only, in speech only, or in both speech and gesture across conditions of congruency and delay |
| Gesture only | Speech only | Speech and gesture | Total |
| Seen congruent | | | |
| Short | 0 | 0 | 0 | 0 |
| Intermediate | 0 | 0 | 0 | 0 |
| Long | 0 | 0 | 0 | 0 |
| Seen incongruent | | | |
| Short | 11 | 0 | 2 | 13 |
| Intermediate | 1 | 0 | 0 | 1 |
| Long | 2 | 4 | 1 | 7 |
| Seen no gesture | | | |
| Short | 0 | 1 | 0 | 1 |
| Intermediate | 0 | 3 | 0 | 3 |
| Long | 1 | 5 | 0 | 6 |
| Total | 15 | 13 | 3 | 31 |
participant seems to have reinterpreted the actor’s gesture as being associated with the character’s reaction to pain. Such a response was coded as a confabulation in speech that was not in the form of a substitution of the target verb.

**DISCUSSION**

Our findings demonstrate that people extract information from gestures when watching others talk, and that over time they rely increasingly on multimodal gist-based representations over representations that retain surface and source information about speech and gesture.

Support for forming multimodal memory representations first comes from our finding that seeing gesture affects gist memory at two levels, which we call “immediate propositional content” and “extended propositional content”. Compared to not seeing a gesture, seeing a gesture increases the likelihood of mentioning the target action (immediate propositional content) and reproducing the target event more completely (extended propositional content). But the semantic relation between speech and gesture (whether it is congruent or incongruent) seems to mainly affect recall for the immediate

| Compatible substitutions | Boomerangs               |
|--------------------------|--------------------------|
|                          | “flies in a boomerang shape” |
|                          | “flies around in a boomerang” |
|                          | “flies in a boomerang kind of thing” |
|                          | “swoops around” |
| Sinks                    | “starts to go underwater” |
|                          | “goes under the surface of the water” |
|                          | “begins to fall before the surface of the water” |
|                          | “dips down” |

| Incompatible substitutions |
|---------------------------|
| Enters lethargically      |
|                          | “bursts in” |
|                          | “comes in hurriedly” |
|                          | “pops into lethargically” |
|                          | “storms in lethargically” |
| Moves steadily            |
|                          | “walks through quickly” |
|                          | “runs through” |
|                          | “something.. strolls” |
|                          | “shuttles through” |
| Zigzags                  |
|                          | “zigzags around” |
| Zips (jacket)            |
|                          | “buttons” |
| Bounces                  |
|                          | “shakes” |
| Accelerate               |
|                          | “walk” |
propositional content: it did not affect how completely people reproduced
sentences, but when people saw a congruent gesture they were more likely to
recall the target action than when they saw an incongruent gesture. Seeing
a gesture may elaborate spatiomotoric features related to the target action,
making it more accessible and likely to be realised during later recall. When
the spatiomotoric features suggested in speech and in gesture are incompara-
tible, gestures may be less successful at elaborating these features in a gist
representation for the immediate propositional content and thus their benefit
is somewhat reduced. This is consistent with Cassell and colleagues’ (1999)
finding that people are more likely to omit information that had been
accompanied by incongruent gestures than congruent ones.

Although the semantic relationship between speech and gesture affects
recall for the immediate propositional content, it may have a less pervasive
effect on memory for extended propositional content: recall for the target
action was better for congruent than for incongruent gestures, but how
completely the event was recalled did not differ across the two types of
gestures. This could be because the relationship of a gesture to the extended
propositional content is indirect. As we suggested, seeing a congruent gesture
for the target action elaborates the representation for that action more
successfully than an incongruent gesture, and recall for the target action
reflects this, but congruent gestures may not be any more successful than
incongruent ones at elaborating other details of the target event. Since the
scope of both types of gestures is limited to the target action and does not
extend to other semantic features of the target event, both types may
be equally effective at activating representations for other aspects of the
target event. Considering the effect of gestures on the extended propositional
content as indirect could explain the advantage of congruent and incon-
gruent gestures over no gestures: since both congruent and incongruent
gestures elaborate the target action, better access to the target action leads in
both cases to better access to associated details of the event.

While we found that people were more likely to recall the target action
and reproduce the event completely when they had seen an incongruent
gesture than when they hadn’t seen a gesture, previous work found no such
advantage (e.g., Feyereisen, 2006, Experiment 2). However, in Feyereisen
(2006), unlike our study, incongruent gestures could not be integrated with
speech because both their timing and semantic relation to speech were
arbitrary (stimuli were constructed by matching the video of one sentence
with speech from another sentence). It could be argued that rather than
contributing to gist memory, incongruent gestures are remembered better
than those without gestures because they are marked as unusual and
therefore salient. Our survey during debriefing revealed that most partici-
pants did in fact notice speech-gesture incongruities. When asked whether
they noticed anything unusual about the videos, 17 participants mentioned
explicitly the actor’s incongruent gestures and three mentioned broadly the actor’s gestures. Of the remaining 20 participants, upon being debriefed about the congruency manipulation, 15 spontaneously provided an example of an incongruent gesture produced by the actor. But even though most participants noticed at least some instances of incongruent gestures, salience on its own does not account for the overall pattern of results. Specifically, salience does not explain why congruent gestures, which were arguably less salient than incongruent gestures, led to better gist memory for the target action than incongruent gestures. Although salience may partly account for why incongruent gestures generally led to better memory for the target action than no gestures, the advantage of seeing congruent gestures cannot be explained without inferring that people extracted spatiomotoric features from gesture.

Our finding that participants’ memory loss for the target action depended on speech-gesture congruency is in line with the view that processing gesture with speech gives rise to multimodal gist representations. While Church and colleagues (2007) did not find an interaction between the presence of gesture and delay for the correct recall of target statements, our study, by manipulating speech-gesture congruency and focusing on recall for the immediate propositional content of the gesture, did reveal an interaction of speech-gesture congruency and delay. When participants saw target actions accompanied by congruent gestures their recall for the target action was at ceiling over time, whereas when they saw target actions accompanied by incongruent gestures their recall for the target action declined over time. When target actions were not accompanied by gestures, participants’ recall for the target action was relatively low and not affected in a systematic way over time. These patterns lend insight into how gestures contribute to gist representations: they suggest that multimodal memory representations are strongest and most stable when arising from the processing of compatible speech and gestures, are relatively strong but less stable when arising from the processing of incompatible speech and gestures, and are the weakest when arising from processing speech on its own.

Note that delay did not have the same effect on memory for extended propositional content as it did for the immediate propositional content. In fact, participants remembered target events more completely after longer delays than after a short delay (see Figure 2b). As we have noted, this pattern can be explained in terms of hypermnesia, the finding that people recall more about an event with each repeated attempt. Because participants had to recall information from each story three times during the experiment, they may have activated events prompted after an intermediate or long delay during earlier prompting, leading to more complete reproduction compared to events prompted after a short delay. As a result, memory for an event as a whole may increase over time through repeated recall.
Participants’ gestures for the target action also provide support for multimodal gist-based representations arising from processing speech with gesture, to the extent that participants’ gestures reflect a more elaborated representation for the target action. Parallel to the way participants encoded the target action in speech, participants were more likely to encode the target action in gesture when they had seen a gesture than when they had not, and more so when the target action was accompanied by a congruent than by an incongruent gesture (see Figure 3a). Moreover, producing a gesture for the target action was more likely after a short delay than after longer delays, particularly for events with incongruent gestures.

The distribution of incongruent gestures, in particular (see Figure 3c and Table 1), is in line with the view that surface information about the actor’s gestures becomes less accessible over time, with people increasingly accessing multimodal gist representations of the events. Because of difficulties during encoding in incorporating spatiomotoric features from incongruent gestures with the accompanying speech, surface information about spatiomotoric features from these gestures is available only briefly. This is similar to findings that people ignore schema-incongruent information when they have an attribution for it or deem it irrelevant, and otherwise accommodate it with their schema (Crocker, Hannah, & Weber, 1983). That is, the difficulty of incorporating incongruent information from gesture with speech led both to the short-lived retention of the surface of incongruent gestures, evidenced by the decrease over time in representational gestures in general and incongruent gestures in particular, and to weaker gist representations over time, evidenced by a decrease in accurately recalling the target action. In fact, over time gist representations of events with incongruent gesture were not only weaker, but also more likely to include inaccurate information: over time participants were more likely to produce inaccuracies in speech stemming from the speech-gesture incongruity. In other words, at longer delays participants were more likely to use a verb or description whose semantic features included the incongruent feature of a gesture (e.g., saying “zigzags around” after seeing a “spiral” gesture and hearing “zigzags”).

Consistent with earlier work showing that gist information is retained better than surface information (e.g., Brewer & Nakamura, 1984; Hjelmquist & Gidlund, 1985; Kintch et al., 1990; Sachs, 1967), participants reproduced target events completely at higher rates than they reproduced them verbatim, despite our instructions to reproduce the actor’s sentences verbatim. However, in our study delay did not have a consistent effect on verbatim memory for the target event. Although we found that people were more likely to reproduce events verbatim when they had seen gestures than when they hadn’t, we did not find that they were better at remembering the surface form of events after short than longer delays. That verbatim recall is better when having seen gestures may initially seem at odds with the finding that recognition of verbatim
sentences is better for statements without gestures than with gestures (e.g., Cutica & Bucciarelli, 2008). However, this difference can be explained by how the measurement tasks access representations elaborated by gesture. In a recognition task like Cutica and Bucciarelli’s (2008), participants were more likely to false alarm on a paraphrase probe of a sentence that was accompanied by gesture because it resonated with more gist-derived information in the representation. In a cued-recall task like ours, participants getting a prompt for a sentence with a gesture were more likely to reproduce surface form aspects of the sentence because they accessed both more gist (paraphrasing) and verbatim information. This is consistent with fuzzy-trace theory’s proposal that remembering is supported by both verbatim and gist traces, with the retrieval of verbatim traces supporting a more vivid form of remembering (recollection) and of gist traces supporting a more generic form of remembering (familiarity) (Brainerd & Reyna, 2002). Despite our participants’ verbatim recall not showing a consistent effect of delay, their decreasing incongruent gestures and their increasing inaccuracies in speech over time do suggest that surface information is, overall, not well preserved.

Overall, our participants’ recall of the target action and extended event, their distribution of gestures, and their inaccuracies suggest that people over time rely on gist-based mental models arising from the processing gestures along with speech, and lose information about the source modality and its surface form. The spatiomotoric features of gesture elaborate memory representations for both the immediate propositional content co-occurring with gesture and the extended propositional content, indirectly. Multimodal memory representations are strongest and most stable over time when the spatiomotoric features of gesture are congruent with those conveyed in speech. When they are incongruent, information originally presented in gesture becomes, over time, less likely to be encoded in gesture and more likely to be encoded in speech, since representations with surface information about the two modalities deteriorate rapidly and people rely instead on multimodal gist representations. Over time, the distinction between information originally carried by speech and information originally carried by gesture diminishes, and memory is dominated by multimodal gist representations.

References

Alibali, M., Flevares, L., & Goldin-Meadow, S. (1997). Assessing knowledge conveyed in gesture: Do teachers have the upper hand? Journal of Educational Psychology, 89, 183–193.
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*, 169–188.

Anderson, J. R., & Bower, G. H. (1973). *Human associative memory*. Washington, DC: Winston.

Bavelas, J. B., & Chovil, N. (2000). Visible acts of meaning: An integrated message model of language in face-to-face dialogue. *Journal of Language and Social Psychology, 19*(2), 163–194.

Beattie, G., & Shovelton, H. (1999). Mapping the range of information contained in the iconic gestures that accompany spontaneous speech. *Journal of Language and Social Psychology, 18*, 438–462.

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

Brainerd, C. J., & Reyna, V. F. (2002). Fuzzy-trace theory and false memory. *Current Directions in Psychological Science, 11*(5), 164–169.

Brainerd, C. J., Reyna, V. F., & Brandse, E. (1995). Are children’s false memories more persistent than their true memories? *Psychological Science, 6*, 359–364.

Brewer, C. J., & Nakamura, G. V. (1984). The nature and functions of schemas. In R. S. Wyer Jr. & T. K. Srull (Eds.), *Handbook of social cognition* (pp. 119–160). Hillsdale, NJ: Lawrence Erlbaum.

Bugental, D. E., Kaswan, J. W., & Love, L. R. (1970). Perception of contradictory meanings conveyed by verbal and nonverbal channels. *Journal of Personality and Social Psychology, 16*, 647–655.

Carney, R. N., & Levin, J. R. (2002). Pictorial illustrations still improve students’ learning from text. *Educational Psychology Review, 13*(1), 5–26.

Cassell, J., McNeill, D., & McCullough, K-E. (1999). Speech-gesture mismatches: Evidence for one underlying representation of linguistic and nonlinguistic information. *Pragmatics and Cognition, 7*(1), 1–33.

Church, R. B., Garber, P., & Rogalski, K. (2007). The role of gesture in memory and social communication. *Gesture, 7*(2), 137–158.

Church, R. B., & Golden-Meadow, S. (1986). The mismatch between gesture and speech as an index of transitional knowledge. *Cognition, 23*, 43–71.

Crocker, J., Hannah, D., & Weber, R. (1983). Person memory and causal attributions. *Journal of Personality and Social Psychology, 44*, 55–66.

Cutica, I., & Buccarelli, M. (2008). The deep versus the shallow: Effects of co-speech gesture in learning from discourse. *Cognitive Science, 32*, 921–935.

Feyereisen, P. (2006). Further investigation on the mnemonic effect of gestures: Their meaning matters. *European Journal of Cognitive Psychology, 18*(2), 185–205.

Goldin-Meadow, S., Nusbaum, H., Garber, P., & Church, R. B. (1993). Transitions in learning: Evidence for simultaneously activated strategies. *Journal of Experimental Psychology: Human Perception and Performance, 19*(1), 92–107.

Goldin-Meadow, S., & Singer, M. A. (2003). From children’s hands to adults’ ears: Gesture’s role in teaching and learning. *Developmental Psychology, 39*(3), 509–520.

Goldin-Meadow, S., Wein, D., & Chang, C. (1992). Assessing knowledge through gesture: Using children’s hands to read their minds. *Cognition and Instruction, 8*(3), 201–209.

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.

Grice, P. H. (1967/1989). *Logic and conversation, studies in the way of words*. Cambridge, MA: Harvard University Press.

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.

Hjelmquist, E., & Gidlund, A. (1985). Free recall of conversations. *Text, 5*(3), 169–185.

Holle, H., & Gunter, T. C. (2007). The role of iconic gestures in speech disambiguation: ERP evidence. *Journal of Cognitive Neuroscience, 19*(7), 1175–1192.
APPENDIX 1

Target events are sentences 4, 9, and 14. Target verbs in these sentences are indicated in bold. Sentences in italics served as the prompts for the target sentence following them.

The ski trip

1. Melissa and her three friends are going on a ski trip.
2. When everyone has boarded the car, Melissa starts the engine and begins navigating the usual route to the mountains.
3. As they’re driving up a steep slope, a large rock falls from the side of the mountain.
4. The car swerves to the side of the road to avoid it.
5. Thankfully, the car just misses the rock, and everyone takes a few minutes to recover from the shock.
6. Melissa gets back on the road and continues driving very carefully.
7. After an hour of uneventful driving, they finally reach the popular ski resort.
8. Melissa parks the car, and everyone unloads their skiing gear and starts dressing up.
9. Melissa zips up her jacket, and puts on her hat, gloves, and sunglasses.
10. Since they are all expert skiers, they decide to go straight to the challenging slopes.
11. So, after warming up, they take the lift to the black diamond slope.
12. Melissa gazes down the steep slope and takes a deep breath.
13. She contemplates her path and when she’s ready she digs her poles in the snow and pushes forward.
14. Melissa zigzags down the slope and safely reaches the foot of the mountain.
15. After an entire afternoon of skiing, they all go back to the cabin to get some warm food and hot cocoa.

Congruent: swerve
Incongruent: zigzag {accompanying gesture has incongruent path of motion: spiral}
No-gesture: zip