Psychon Bull Rev (2015) 22:517–523
DOI 10.3758/s13423-014-0681-7

BRIEF REPORT

The processing of speech, gesture, and action during language comprehension

Spencer Kelly · Meghan Healey · Asli Özyürek · Judith Holler

Published online: 8 July 2014

Abstract Hand gestures and speech form a single integrated system of meaning during language comprehension, but is gesture processed with speech in a unique fashion? We had subjects watch multimodal videos that presented auditory (words) and visual (gestures and actions on objects) information. Half of the subjects related the *audio* information to a written prime presented before the video, and the other half related the *visual* information to the written prime. For half of the multimodal video stimuli, the audio and visual information contents were congruent, and for the other half, they were incongruent. For all subjects, stimuli in which the gestures and actions were incongruent with the speech produced more errors and longer response times than did stimuli that were congruent, but this effect was less prominent for speech–action stimuli than for speech–gesture stimuli. However, subjects focusing on visual targets were more accurate when processing actions than gestures. These results suggest that although actions may be easier to process than gestures, gestures may be more tightly tied to the processing of accompanying speech.

Keywords Language comprehension · Embodied cognition · Gesture · Action

S. Kelly (✉) · M. Healey
Department of Psychology, Neuroscience Program, Colgate University, 105D Olin Hall, Hamilton, NY, USA
e-mail: skelly@colgate.edu

S. Kelly
Center for Language and Brain, Colgate University, Hamilton, NY, USA

A. Özyürek
Radboud University, Nijmegen, The Netherlands

A. Özyürek · J. Holler
Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands

A. Özyürek
Donders Institute for Brain, Cognition, and Behavior, Nijmegen, The Netherlands

M. Healey
Neuroscience Graduate Group, University of Pennsylvania, Philadelphia, PA, USA

Gesture and speech are theorized to form a single integrated system of meaning during language production (Kendon 1986; McNeill 1992), but whereas speech provides information sequentially through arbitrary symbols, gesture is holistic and imagistic. Together, they create a more complete message than either modality can alone (Clark 1996).

Similarly, it seems that a speaker's message cannot be fully *understood* without attention to both modalities. The recently proposed "integrated-systems hypothesis" posits that speech and gesture are tightly integrated and mutually and obligatorily interact in order to enhance language comprehension (Kelly, Özyürek, and Maris 2010). This hypothesis helps to explain numerous behavioral studies showing that gestures play a significant role in comprehension (for a review and meta-analysis, see Hostetter 2011).

It has not been clear from the above research, however, whether gesture is *uniquely* connected to speech during language comprehension. Some neuroimaging evidence suggests that the brain comprehends gestures in a similar fashion to other sorts of visual information, such as pictures (Willems, Özyürek, and Hagoort 2008; Wu and Coulson 2011) and pantomimes (Willems, Özyürek, and Hagoort 2009). In the present study, we extended this research by focusing on a more pervasive and visually rich and dynamic type of cospeech information: manual actions on objects. Manual actions, especially when they are used communicatively, are one of the most ubiquitous and commonly occurring forms of visual input during social interactions (for more on actions on objects, see Rosenbaum et al. 2012). But it is

not well understood how comprehension of actions relates to the comprehension of hand gestures, which are equally pervasive and natural. For example, imagine the spoken utterance "This is how you tie a slip knot," accompanied by two types of visual input: either a gesture *depicting* tying the knot, or actually *tying the knot*. Which illustration is more communicative?

It is possible that gestures and actions have similar communicative power.[1] Indeed, Hostetter and Alibali (2008) argued that gestures are simulations of actual actions and, consequently, may be part of a shared system in language production (see also Hostetter and Alibali 2010; Wagner Cook and Tanenhaus 2009).

Then again, gestures are much more abstract and less visually "complete" than actions, so perhaps they are a lesser substitute for the real thing. According to Paivio's (1986) dual-coding theory, the more concrete and imagistic the visual code, the more easily it will be understood and affect processing of the accompanying spoken code. Because actions differ from gestures in that actual objects are physically present with the actions, they may provide the listener-viewer with a richer and more concrete imagistic representation of the speech content than do gestures. In this way, actions may be easier to process and have a larger effect on speech.

However, there are compelling reasons to believe that gestures may be at least as communicative as—if not more communicative than—actions. People have extensive experience producing and observing actions that are not intended to communicate, but instead, are designed to accomplish nonsocial, and more instrumental, goals. Indeed, briefly consider the myriad of actions you have done or seen today that fall into this category (e.g., drinking coffee, typing on a keyboard, answering the phone, etc.). In contrast, most spontaneous hand gestures—conventional signs and emblems excluded—are seldom produced in the absence of speech or an interlocutor. For example, imagine chopping vegetables for dinner while simultaneously talking about your day at work. In this scenario, it is easy to picture how your interlocutor would not view your actions as being meant to go with your speech. In contrast, it would be perplexing if you talked about your day at work while simultaneously producing chopping gestures (unless, of course, you are a chef). From this perspective, gestures might actually have more communicative power than actions, making them more tightly integrated with the accompanying speech (Kendon 2004; McNeill 2012).

In the present research, we explored this issue using a paradigm similar to that of Kelly et al. (2010). Subjects were presented with multimodal stimuli in which speech was congruent or incongruent with either a gesture or action. These stimuli were preceded by a written prime, and the task was to determine whether either the audio or the visual information (gesture/action) in the multimodal stimuli was related to the written prime. To investigate the relative communicative strengths of gesture and action, we were interested in two measures of processing: (1) the overall speed and accuracy of identifying the two types of visual targets (gestures and actions), and (2) the extent to which incongruities between speech and gesture or speech and action disrupted processing. The extent to which participants cannot ignore irrelevant information within gesture–speech and action–speech stimuli is a measure of the relative strength of their integration (see Kelly et al. 2010, for more on this measure of multimodal integration).

We had two goals for this study. The first was to replicate previous research and show that incongruent multimodal utterances are harder to process than congruent onces. The second was to test two competing predictions: If the rich imagery and concreteness of actions make them a more powerful source of visual information than gesture, subjects should (1) process the action targets more easily and (2) have more difficulty ignoring the irrelevant information in action–speech videos than in gesture–speech videos. In contrast, if gestures are privileged and have a special relationship with speech, subjects should show the opposite pattern.

Method

Subjects

A group of 62 right-handed, native English-speaking college undergraduates between the ages of 18 and 22 years (39 females, 23 males) participated.[2] The university's Institutional Review Board approved the experiment, and all subjects gave written informed consent prior to participation.

Materials

Digitized videos were created with a Sony DV100 digital camcorder and edited with Final Cut Pro software. The videos contained a female actress situated in natural contexts (e.g., kitchen, living room, entry way, etc.) describing everyday activities (e.g., drinking coffee, watering plants, tying shoes, etc.). The actress faced the camera in all videos, but her face was digitally covered to block access to lip information (see below for why we dubbed the speech). Subjects were told that this was to hide the actress's identity, to avoid distractions for people who knew her. The actress spoke at a regular speaking rate with no artificial pauses between the words.

[1] In the present article, we will simply use the term "actions" to refer to "manual actions on objects."

[2] This sample size was chosen because it is similar to that in previous research using a very similar design (Kelly et al. 2010).

The spoken utterance always followed a simple verb–object pattern, spoken in the past tense (e.g., "Drank the coffee," "Filed nails," "Tied the shoes"). The speech was recorded offline and carefully inserted into the videos, so that it was identical across all of the experimental conditions (Fig. 1). Equating the audio across all conditions was important, to control for subtle differences in how producing gestures and actions might alter the acoustic properties of the accompanying speech (Krahmer and Swerts 2007). Because the lips were digitally obscured, it was not a problem to temporally align the speech with the visual gestures and actions.

A total of 30 spoken sentences were created. Each sentence was produced in four different conditions—congruent gesture, congruent action, incongruent gesture, and incongruent action—for a total of 120 multimodal clips. In the congruent gesture condition, the cospeech gestures provided a visual representation of the speech. For example, accompanying the sentence "Poured water" was a gesture of pouring. There was no subject in the sentence, because we did not want any additional linguistic information to affect the processing of the verbs, which served as the onset of our response time (RT) measure. The congruent action condition was identical, except actual objects were present (e.g., the actor poured water from a pitcher to a glass). To create the incongruent conditions, the speech was paired with gestures/actions from a different congruent pairing. For example, the incongruent gesture/action for "Poured water" was gesturing or actually chopping vegetables. See Fig. 1 for examples of the congruent and incongruent videos.

The actress was asked to perform the gesture/action stroke (i.e., the most meaningful part of the movement, such as the pouring component, rather than the preparatory movement leading up to it; McNeill 1992) with the speech naturally when filming the vignettes. Great care was taken to ensure that the only difference between the action and gesture videos was the presence or absence of objects. Each multimodal stimulus began with the stroke of the gesture or the corresponding onset of the action. As in Kelly et al. (2010), the audio component was inserted in the video 200 ms after the onset of the visual component of the gesture or action.

Prior to each stimulus, a written word was displayed on the screen, to serve as the prime in the task (see below). In previous research employing a similar paradigm (Kelly et al. 2010), we had used videos of actions as the primes, but because we were comparing actions to gestures in the present study, a more neutral type of prime was necessary. The word was displayed for 500 ms, followed by a blank screen for 500 ms prior to stimulus onset. Each word displayed was a mannered verb used in one of the experimental sentences, and it was either related or unrelated to the auditory and/or the visual information presented in the video. The variable intertrial interval between each prime–target pair ranged from 1 to 2.5 s.

Procedure

The main task was for subjects to relate the written word (prime) to either the audio (half of the subjects) or the visual (the other half of the subjects) portion of the video stimuli. In addition, to ensure that subjects focused on both the auditory and visual modalities, they were told that they would take a recall task about the auditory and visual aspects of the videos at the conclusion of the experimental session. This deception was meant to distract the subjects from the true experimental purpose (to compare the processing of gestures and actions) and to ensure that they paid attention to both modalities during the experiment. No recall task was actually administered.

Subjects viewed a randomized sequence of the 120 short video clips twice, once with a related prime and once with an unrelated prime (never in succession), for a total of 240 stimulus presentations. Half of the subjects (N=31) were instructed to focus on the auditory information in the stimuli: to press one button if the *speech* of the actress related to the written prime, and to press another button if the speech it did not relate. The other half of the subjects (N=31) were instructed to focus on the visual information in the stimuli: to press one button if the *visual content* (gesture or action) of the speaker's movements related to the written prime, and to press another button if it did not relate. The stimuli were constructed across the two groups, such that half of the trials were "related" and the other half were "unrelated." For example, referring to Fig. 1, suppose that the prime was the word "Pour." For the speech target task, Panels A and C are related, and Panel B is unrelated. For the visual movement target task, Panels A and B are related, and Panel C is unrelated. In addition, one other type of stimulus (unrelated and incongruent) was presented, to create a completely balanced design. The experimental procedure lasted approximately 30 min. Refer to Table 1.

Design and analysis

The design was mixed, with Target Type (audio, visual) as a between-groups factor and Movement Type (gesture, action), Audiovisual Congruence (congruent, incongruent), and Priming Relationship (related, unrelated) as within-subjects factors.[3] Separate four-way analyses of variance were run on the error rate and RT data, with planned orthogonal t tests (one-tailed) following up on the Movement Type × Audiovisual Congruence interaction (which was the primary interaction of interest). To conserve space, only statistically significant effects are reported.

[3] Although we did not make predictions about differences between the "related" and "unrelated" trials, we included both conditions in the analysis to increase power. A subanalysis focusing on just the "related" trials yielded a similar, but less robust, pattern with respect to the fuller one reported here.

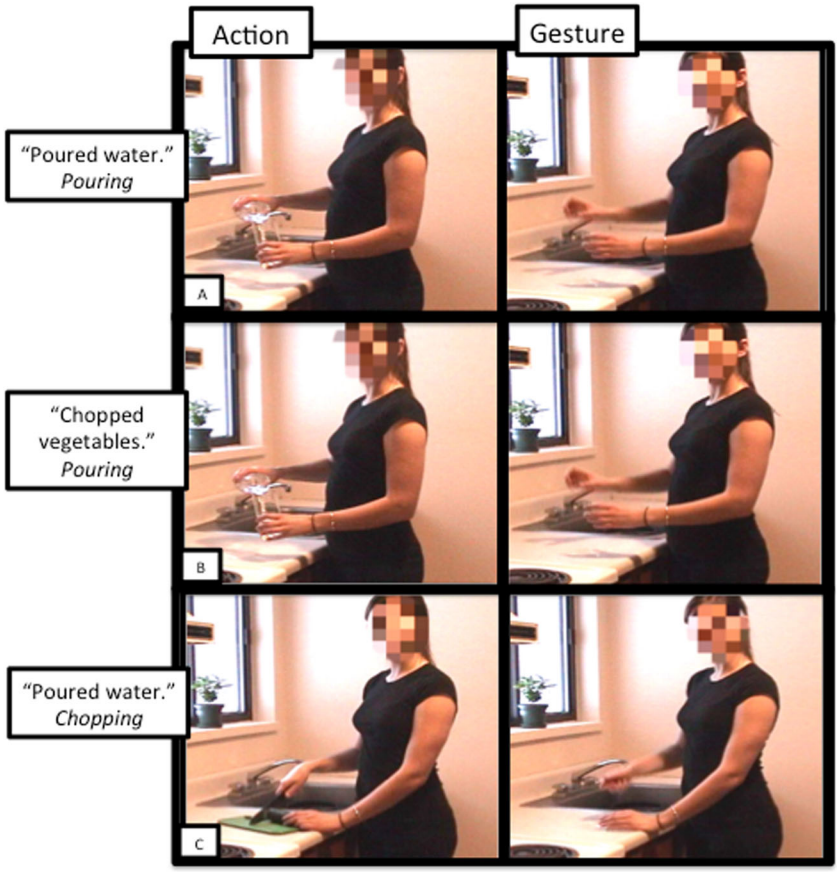

Fig. 1 Actions are on the left, gestures on the right. In each panel, the speech is in quotes and the visual information is in italics. On the basis of these two pieces of information in each panel, Panel A is a congruent stimulus, and Panels B and C are incongruent

Results

As a précis of the most important results of the primary analyses, subjects in both the audio- and visual-target tasks were slower and less accurate when processing stimuli in which the gestures and actions were incongruent versus congruent with the speech, but this effect was less prominent for speech–action than for speech–gesture stimuli. However, subjects in the visual-target task were faster and more accurate when identifying actions versus gestures.

Error rates

We found a significant main effect of audiovisual congruence, $F_1(1, 60) = 22.90, p < .001, \eta_p^2 = .28$ (by subjects); $F_2(1, 56) = 7.15, p = .01, \eta_p^2 = .11$ (by items), with subjects making fewer errors when identifying (audio and visual) targets in congruent (M = 2.7%) than in incongruent (M = 4.0%) stimuli. In addition, a significant main effect of movement type was visible, $F_1(1, 60) = 8.03, p = .006, \eta_p^2 = .12$; $F_2(1, 56) = 0.62$, n.s., with actions (M = 2.8%) producing fewer errors than did

Table 1 Examples of the stimuli in the congruent and incongruent conditions for related and unrelated primes within each audio- and visual-target task

Prime		Congruent Audio	Congruent Visual	Incongruent Audio	Incongruent Visual
Audio Targets					
Related:	*Scrubbed*	"Scrubbed the dishes"	Scrubbing	"Scrubbed the dishes"	Chopping
Unrelated:	*Calculated*	"Scrubbed the dishes"	Scrubbing	"Scrubbed the dishes"	Chopping
Visual Targets					
Related:	*Scrubbed*	"Scrubbed the dishes"	Scrubbing	"Chopped the vegetables"	Scrubbing
Unrelated:	*Calculated*	"Scrubbed the dishes"	Scrubbing	"Chopped the vegetables"	Scrubbing

Note that all subjects received two sets of these stimuli, one with gestures and one with actions as the visual information.

gestures (M = 3.9%), as well as an effect of prime relationship, $F_1(1, 60) = 34.13$, $p < .001$, $\eta_p^2 = .36$; $F_2(1, 56) = 26.94$, $p < .001$, $\eta_p^2 = .33$, with related trials (M = 4.4%) producing more errors than did unrelated trials (M = 2.3%). We observed no significant main effect of target type, $F_1(1, 60) = 2.33$, n.s., for subjects, but there was a significant effect for items, $F_2(1, 56) = 13.66$, $p < .001$, $\eta_p^2 = .20$, with subjects making more errors with visual (M = 3.9%) than with audio (M = 2.8%) targets.

A significant Movement Type×Target Type interaction emerged, $F_1(1, 60) = 10.28$, $p < .001$, $\eta_p^2 = .15$; $F_2(1, 56) = 6.05$, $p = .017$, $\eta_p^2 = .10$, such that when the target type was auditory, the gesture and action conditions did not differ, $t_1(30) = 0.86$, n.s.; $t_2(28) = 1.02$, n.s., but when the target was visual, gestures (M = 4.7%) produced significantly more errors than did actions (M = 3.1%), $t_1(30) = 4.07$, $p < .001$; $t_2(28) = 2.90$, $p = .007$. In addition, a significant Audiovisual Congruence×Prime Relationship interaction was apparent, $F_1(1, 60) = 64.37$, $p < .001$, $\eta_p^2 = .52$; $F_2(1, 56) = 81.94$, $p < .001$, $\eta_p^2 = .60$, with congruent stimuli producing fewer errors (M = 2.2%) than incongruent stimuli (M = 6.6%) in the related condition, $t_1(61) = 8.81$, $p < .001$; $t_2(57) = 7.49$, $p < .001$, but incongruent stimuli (M = 1.4%) producing fewer errors than congruent stimuli (M = 3.2%) in the unrelated condition, $t_1(61) = 3.79$, $p < .001$; $t_2(57) = 4.83$, $p = .007$.

Focusing on the primary analysis of interest, we observed a significant Movement Type×Audiovisual Congruence interaction, $F_1(1, 60) = 4.50$, $p = .038$, $\eta_p^2 = .07$; $F_2(1, 56) = 2.88$, $p = .10$, $\eta_p^2 = .05$, which was driven by the gesture condition: Congruent gesture–speech stimuli produced fewer errors (M = 2.6%) than did incongruent stimuli (M = 4.8%), $t_1(61) = 5.38$, $p < .001$; $t_2(57) = 3.56$, $p < .001$, whereas in the action condition, no differences were visible between congruent and incongruent action–speech stimuli, $t_1(61) = 1.34$, n.s.; $t_2(57) = 0.73$, n.s. Refer to Fig. 2.

Response times

We found significant main effects of audiovisual congruence, $F_1(1, 60) = 6.76$, $p = .012$, $\eta_p^2 = .10$; $F_2(1, 56) = 6.33$, $p = .015$, $\eta_p^2 = .10$, with subjects responding faster to targets (audio and visual) in congruent (M = 1,262 ms) than in incongruent (M = 1,282 ms) stimuli, and of prime relationship, $F_1(1, 60) = 16.43$, $p < .001$, $\eta_p^2 = .22$; $F_2(1, 56) = 18.30$, $p < .001$, $\eta_p^2 = .25$, with related trials (M = 1,245 ms) producing faster RTs than unrelated trials (M = 1,299 ms). No significant main effect of target type emerged by subjects, $F_1(1, 60) = 2.44$, n.s., but we did observe a significant effect by items, $F_2(1, 56) = 59.74$, $p < .001$, $\eta_p^2 = .52$, with subjects being faster with visual (M = 1,196 ms) than with audio (M = 1,347 ms) targets.

A significant Movement Type×Target Type interaction was apparent by subjects, $F_1(1, 60) = 6.41$, $p = .014$, $\eta_p^2 = .1$, but not by items, $F_2(1, 56) = 0.80$, n.s., such that when the target type was auditory, the gesture and action conditions did not differ, $t_1(30) = 1.00$, n.s.; $t_2(28) = 0.72$, n.s., but when the target was visual, gestures (M = 1,212 ms) produced slower RTs than did actions (M = 1,181 ms), $t_1(28) = 2.41$, $p = .022$; $t_2(28) = 1.50$, n.s.

For the primary analysis, despite revealing no significant Movement Type×Audiovisual Congruence interaction, $F_1(1, 60) = 0.85$, n.s.; $F_2(1, 56) = 0.38$, n.s., our planned t tests did show that within the gesture condition, subjects were faster to identify targets for congruent stimuli (M = 1,265 ms) than for incongruent stimuli (M = 1,290 ms), $t_1(61) = 2.69$, $p = .005$; $t_2(57) = 1.75$, $p = .043$, but the action condition produced less definitive results, with congruent stimuli (M = 1,259 ms) being significantly faster than incongruent stimuli (M = 1,274 ms) only for items, $t_2(57) = 2.19$, $p = .017$, but not for subjects, $t_1(61) = 1.52$, n.s.. See Fig. 3.

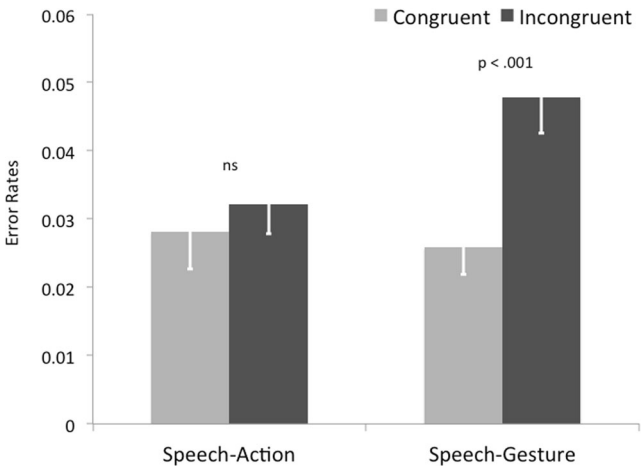

Fig. 2 Error rates (as proportions, with standard errors) for congruent and incongruent speech–action and speech–gesture stimuli, collapsed across the audio- and visual-target tasks. Significance is shown by subjects

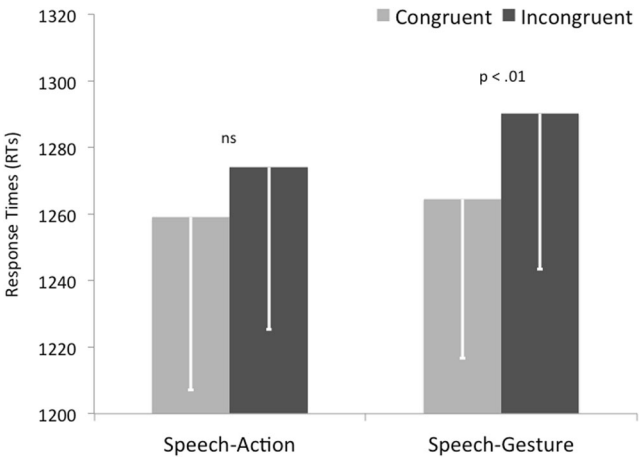

Fig. 3 Response times (with standard errors) for congruent and incongruent speech–action and speech–gesture stimuli, collapsed across the audio- and visual-target tasks. Significance is shown by subjects

Discussion

The results achieved our first goal, clearly demonstrating that incongruent stimuli in both the gesture and action conditions produced more errors and longer RTs than did congruent stimuli for subjects in both the audio- and visual-target conditions. We also found interesting results for our second prediction: Whereas subjects identified action targets more accurately (and were also faster when the target was visual), incongruent gestures were more disruptive to processing speech than incongruent actions.

The incongruency effect replicates and extends previous work by Kelly et al. (2010). In the present study, we found not only a bidirectional influence of gesture and speech, but also one of action and speech, such that when the task was to focus on auditory speech, performance suffered when the speech was accompanied by incongruent gestures and actions, and when the task was to focus on gestures and actions, performance suffered when they were accompanied by incongruent auditory speech.[4] Moreover, as in previous work, this integration appears hard to control: Even when one modality was completely irrelevant to the task—be it auditory language or visual gesture and action—subjects could not ignore information conveyed in that modality.

However, the results from our second prediction tell a more nuanced story. On one hand, actions appear to be easier to process than gestures, a finding that makes sense from the perspective of Paivio (1986), who predicted that the richer the visual code, the easier it would be to understand. Indeed, the speech–action stimuli were more concrete and visually complete, and this may have made things easier for subjects when they did the priming task, particularly for subjects in the visual-target task. In this way, one may conclude that actions, at a basic perceptual level, are visually more informative than gestures. On the other hand, this advantage of actions makes the second part of the prediction—that is, that gesture–speech disruptions would be greater than action–speech disruptions—even more interesting. These combined results suggest that even though gestures are visually less informative than actions, they may be treated as *communicatively* more informative in relation to the accompanying speech. In other words, although gestures are stripped of much of the visual richness of actions, something important remains.

One possibility of what remains has less to do with what is (or is not) in the hands of the gesturer, and more to do with what is in the heads of the viewer: Viewers may generally assume that gesture, more than action, is information *meant to* accompany speech, and this may increase their attention to it. Indeed, a gesture, like a word, is always a representation of something else, whereas an action on an object need not mean anything more than accomplishing an instrumental goal for the actor. This abstractness of gestures is very conducive to communication. For example, if a friend were to tell you that a group of groggy coworkers "were up late last night," while (1) simultaneously producing a drinking gesture or (2) drinking from an actual beverage, you would interpret the gesture as clarifying the speech (i.e., *why* your coworkers were up late), whereas you might view the actual act of drinking as either clarifying the speech or simply quenching your friend's immediate thirst. In other words, people may view gestures as being more intended than actions to accompany speech, and this pragmatic awareness may impact the semantic processing of that speech.

This possibility receives support from previous work that has attempted to manipulate the communicative intent behind hand gestures (Holler et al. 2014; Kelly, Ward, Creigh, and Bartolotti 2007). For example, Kelly et al. (2007)) told people that certain gestures were not intended to accompany speech, and they found a reduction in the neural integration of those particular gesture–speech pairs during comprehension. More recently, Holler et al. (2014)) reduced the communicative intent associated with gestures by averting a speaker's (gesturer's) eye gaze away from subjects (thus rendering them unaddressed recipients) and found that multimodal processing areas in the brain (right middle temporal gyrus) differentiated utterances conveyed through speech versus speech-plus-gesture to a lesser extent than when the subjects were direct addressees.

Of course, this same sort of intentional communication is possible with actions. After all, people often successfully use actions on objects to communicate important information to interlocutors (e.g., a chef showing an apprentice how to safely prepare blowfish), and these actions no doubt add crucial meaning for the addressees (Clark 1996). In fact, in typical face-to-face interactions, speakers can use all sorts of cues (e.g., deictic terms, holding up objects, gazing between object and addressee, etc.) to indicate that an action is relevant to what one is saying. What we are suggesting is that hand gestures may simply need fewer of these contextual cues to effectively communicate their relevance to speech. In this way, it is worth thinking of gestures as specifically *designed for* communication, whereas actions on objects can be *co-opted* for it, but only if the context is right. In fact, we believe that the present study may have inflated the communicative power of the actions on objects, because subjects likely viewed most of them as being communicatively intended (why else would an experimenter be showing them these videos, if the actions were not relevant?), so it will be important for future research to compare the processing of cospeech gestures and actions in more natural, face-to-face social interactions.

[4] Although we do not report them here, we ran contrasts on subjects within both the audio- and visual-target conditions ($N = 31$ in each condition), and error rates and RTs showed the same pattern of results.

Before we conclude, it is necessary to note that our most central finding, the Movement Type×Audiovisual Congruence interaction, was not very robust. However, the congruence main effect was quite solid, and that finding—even by itself—still tells an interesting story: Despite the fact that gestures are visually less complete and concrete than actions, they were just as strongly integrated with speech. So, whereas gestures may be relatively "impoverished" in visual imagery, as compared to actions, they are rich in their own way, and may actually do more with less (for more on the special representational richness of gesture in problem solving, see Novack, Congdon, Hemani-Lopez, and Goldin-Meadow 2014).

In conclusion, the findings that gestures and communicative actions are equally hard to ignore and bidirectionally interact with speech are in keeping with research suggesting that gestures are part of the same cognitive system as actions (Hostetter and Alibali 2008, 2010; Wagner Cook and Tanenhaus 2009). However, the fact that actions are processed faster and more accurately than gestures, but that gestures may influence speech to a greater extent than actions, suggests a more nuanced picture. Although actions are informationally richer and more impactful than gestures (Paivio 1986), the communicative relationship between speech and gesture may be closer than that between speech and action (McNeill 2012), and this may give gestures and actions different statuses in language comprehension.

Author note The authors thank Sasha Ivanov for being the actress in the stimulus videos; Lauren Okada for her assistance in filming; and Emily Borden, Emma Krasovich, Jessica Halter, Carmen Lin, Raman Malik, and Andrew Wylie for their assistance with collecting data. J.H. was supported through a Marie Curie Fellowship (255569) and European Research Council Advanved Grant INTERACT (269484).

References

Clark, H. H. (1996). *Using language*. Cambridge, UK: Cambridge University Press.

Holler, J., Kokal, I., Toni, I., Hagoort, P., Kelly, S. D., & Özyürek, A. (2014). Eye'm talking to you: Speakers' gaze direction modulates co-speech gesture processing in the right MTG. *Social Cognitive and Affective Neuroscience*. doi:10.1093/scan/nsu047. Advance online publication.

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

Hostetter, A. B., & Alibali, M. W. (2008). Visible embodiment: Gestures as simulated action. *Psychonomic Bulletin & Review, 15,* 495–514. doi:10.3758/PBR.15.3.495

Hostetter, A. B., & Alibali, M. W. (2010). Language, gesture, action! A test of the Gesture as Simulated Action framework. *Journal of Memory and Language, 63,* 245–257.

Kelly, S. D., Ward, S., Creigh, P., & Bartolotti, J. (2007). An intentional stance modulates the integration of gesture and speech during comprehension. *Brain and Language, 101,* 222–233.

Kelly, S. D., Özyürek, A., & Maris, E. (2010). Two sides of the same coin: Speech and gesture mutually interact to enhance comprehension. *Psychological Science, 21,* 260–267. doi:10.1177/0956797609357327

Kendon, A. (1986). Current issues in the study of gesture. In J.-L. Nespoulous, P. Perron, A. Roch Lecours, & the Toronto Semiotic Circle (Eds.), *The biological foundations of gestures: Motor and semiotic aspects* (Vol. 1, pp. 23–47). Hillsdale, NJ: Erlbaum.

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

Krahmer, E., & Swerts, M. (2007). The effects of visual beats on prosodic prominence: Acoustic analyses, auditory perception and visual perception. *Journal of Memory and Language, 57,* 396–414.

McNeill, D. (1992). *Hand and mind: What gestures reveal about thought*. Chicago, IL: University of Chicago Press.

McNeill, D. (2012). *How language began: Gesture and speech in human evolution*. New York, NY: Cambridge University Press.

Novack, M. A., Congdon, E. L., Hemani-Lopez, N., & Goldin-Meadow, S. (2014). From action to abstraction using the hands to learn math. *Psychological Science, 25,* 903–910.

Paivio, A. (1986). *Mental representations: A dual coding approach*. Oxford, UK: Oxford University Press.

Rosenbaum, D. A., Chapman, K. M., Weigelt, M., Weiss, D. J., & van der Wel, R. (2012). Cognition, action, and object manipulation. *Psychological Bulletin, 138,* 924–946.

Wagner Cook, S., & Tanenhaus, M. K. (2009). Embodied communication: Speakers' gestures affect listeners' actions. *Cognition, 113,* 98–104. doi:10.1016/j.cognition.2009.06.006

Willems, R. M., Özyürek, A., & Hagoort, P. (2008). Seeing and hearing meaning: ERP and fMRI evidence of word versus picture integration into a sentence context. *Journal of Cognitive Neuroscience, 20,* 1235–1249. doi:10.1162/jocn.2008.20085

Willems, R. M., Özyürek, A., & Hagoort, P. (2009). Differential roles for left inferior frontal and superior temporal cortex in multimodal integration of action and language. *NeuroImage, 47,* 1992–2004.

Wu, Y. C., & Coulson, S. (2011). Are depictive gestures like pictures? Commonalities and differences in semantic processing. *Brain and Language, 119,* 184–195. doi:10.1016/j.bandl.2011.07.002