COGNITIVE SCIENCE
A Multidisciplinary Journal

Cognitive Science 39 (2015) 1855–1880

ISSN: 0364-0213 print / 1551-6709 online
DOI: 10.1111/cogs.12221

The Development of the Ability to Semantically Integrate Information in Speech and Iconic Gesture in Comprehension

Kazuki Sekine,[a] Hannah Sowden,[b] Sotaro Kita[a]

[a]*Department of Psychology, University of Warwick*
[b]*Department of English Language and Communication, Kingston University*

Received 18 December 2012; received in revised form 23 October 2014; accepted 27 October 2014

Abstract

We examined whether children's ability to integrate speech and gesture follows the pattern of a broader developmental shift between 3- and 5-year-old children (Ramscar & Gitcho, 2007) regarding the ability to process two pieces of information simultaneously. In Experiment 1, 3-year-olds, 5-year-olds, and adults were presented with either an iconic gesture or a spoken sentence or a combination of the two on a computer screen, and they were instructed to select a photograph that best matched the message. The 3-year-olds did not integrate information in speech and gesture, but 5-year-olds and adults did. In Experiment 2, 3-year-old children were presented with the same speech and gesture as in Experiment 1 that were produced live by an experimenter. When presented live, 3-year-olds could integrate speech and gesture. We concluded that development of the integration ability is a part of the broader developmental shift; however, live-presentation facilitates the nascent integration ability in 3-year-olds.

Keywords: Gesture; Speech; Semantics; Multimodal communication; Preschool children

1. Introduction

In everyday conversation, integrating information from speech and gesture is crucial in computing the speaker's intended message (Cassel, McNeill, & McCullough, 1999; McNeill, 1992). Adults can take into account both gesture and semantically co-expressive words to comprehend an overall message (e.g., Cocks, Sautin, Kita, Morgan, & Zlotowitz, 2009; Goldin-Meadow & Sandhofer, 1999; Kelly, Kravitz, & Hopkins, 2003; Kelly,

Correspondence should be sent to Kazuki Sekine, Department of Psychology, University of Warwick, Coventry CV4 7AL, UK. E-mail: kazuki@tkc.att.ne.jp

Özyürek, & Maris, 2010). The current study focused on the development of this ability in children.

The development of speech–gesture integration may be a part of a larger developmental shift in how children take into account two pieces of information in their behaviors. It has been claimed that a domain-general processing shift occurs between 3- and 5-year-olds (Ramscar & Gitcho, 2007). This is "a shift from behavioral responses driven by a single (often prepotent) factor to those that integrate or select between multiple" factors (Ramscar & Gitcho, 2007, p. 274).

Some researchers have revealed that 5-year-olds can consider two different pieces of information at the same time but not 3-year-olds (e.g., Ramscar & Gitcho, 2007; Zelazo, Muller, Frye, & Marcovitch, 2003). For example, Loewenstein and Gentner (2005; Experiment 3) found that between the age of 3 and 5, children develop the ability to detect an analogy between two representations. Derks and Paclisanu (1967) reported that by the age of 5, children could compare the probabilities of an occurrence of two events and predict which one of the two events was likely to occur next.

Along a similar vein, other researchers have found that 5-year-olds can resolve the conflict between two competing representations, but not 3-year-olds. These findings have been obtained in many fields, such as understanding a false belief (Resches & Perez Pereira, 2007), distinguishing between appearance and reality of an object (Flavell, Green, & Flavell, 1986), sorting card along the different dimensions (Frye, Zelazo, & Palfai, 1995; Zelazo et al., 2003), and inferring a word meaning by the type of predicates (e.g., A predicate "This is made of . . ." implies a material, and a predicate "This has a. . ." implies an object part) (Deák, 2000). As Zelazo et al. (2003) argued, it may be hard for 3-year-olds to process two different representations at the same time, because they tend to perseverate on a salient representation.

The developmental shift, which was proposed by Ramscar and Gitcho (2007), may account for the development of ability to integrate gesture and speech. In this study, *integration* is operationally defined as the listener's derivation of the speaker's message by unifying information from the speaker's gesture and speech, in which the two modalities mutually constrain each other's meanings. For example, when the speaker says, "I was eating a lot" while lifting his fist to his mouth as if holding something, a listener may derive the interpretation that the speaker was bringing food to the mouth with a fork or a spoon, not with his bare hand. If the developmental shift proposed by Ramscar and Gitcho applies to integration of gesture and speech, then 3-year-olds, but not 5-year-olds, should show difficulty in integrating information from gesture and speech, because 3-year-olds have difficulty in processing information from the two modalities at the same time.

Alternatively, it is possible that speech-gesture processing is an exception to Ramscar and Gitcho's (2007) principle, because gesture and speech have special affinity with each other. When speech is presented with an action (e.g., chopping vegetables) or an iconic gesture (e.g., a gesture depicting chopping), the semantic judgment about one modality was influenced by the information in the other modality; more important, this cross-modal influence was stronger for speech-gesture combinations than speech-action combinations

(Kelly, Healey, Özyürek, & Holle, 2014). This finding by Kelly et al. (2014) suggests that considering both gesture and speech is less demanding than considering both action and speech. Thus, even 3-year-olds may be able to integrate information from both gesture and speech.

Though a number of previous studies investigated children's processing of speech-gesture combinations, many of the studies did *not* investigate if children show that the two modalities *mutually constrain* each other's meanings to arrive at a unified message (see the summary in the Table A1 in Appendix A). For example, it has been shown that 5- to 10-year-old children can obtain gestural information when speech and gesture are presented together. For example, 8- and 10-year-olds detected information conveyed solely by iconic gestures when presented with children's explanation of Piagetian conservation tasks (Church, Kelly & Lynch, 2000; Kelly & Church, 1998). When 5- and 6-year-olds responded to interview questions, they incorporated information conveyed solely in an interviewer's iconic gestures (Broaders & Goldin-Meadow, 2010). Another line of studies showed that gestures can constrain the meaning of concurrent speech (but those studies did not show speech-to-gesture constraint). For example, iconic gestures facilitated 1- and 2-year-olds' understanding of the referent of a word that they had poor understanding of (McGregor, Rohlfing, Bean, & Marschner, 2009) and the referent of a novel word (Goodrich & Hudson-Kam, 2009; Mumford & Kita, 2014). In contrast, the current study focused on the question of whether children and adults can integrate speech and iconic gesture that mutually constrain each other's meaning.

In order to investigate this type of integration, it is necessary to compare speech-only, gesture-only, and speech + gesture conditions. Two studies tested children's ability to integrate information from speech and the pointing gesture by comparing all three required conditions (Kelly, 2001; Morford & Goldin-Meadow, 1992). Morford and Goldin-Meadow (1992) compared 14- to 28-month-old children's responses to speech alone sentences (e.g., "push the ball") with responses to speech accompanied by a gesture (e.g., "push the ball" + point at ball or "push" + point at ball). The result showed that the number of correct responses (e.g., they pushed the ball) when presented with the speech and gesture combination was significantly greater than those when presented with the speech alone and those when presented with gesture alone (e.g., pointing at the ball). In Kelly's (2001) study, children aged 3–5 years were asked to interpret an indirect request that an adult made with a pointing gesture and a short sentence (e.g., the adult is saying "it is going to get loud in here" while pointing at the door). The proportion of times the children responded correctly (i.e., they closed the door) when presented with speech alone or with gesture alone was significantly smaller than that when presented with gesture + speech.

The findings from these two studies (Kelly, 2001; Morford & Goldin-Meadow, 1992) suggest that children can integrate speech and the pointing gesture. However, it is not clear from the two studies if children's performance differs from adults as neither studies reported adult data. More specifically, the two studies showed that children performed better when presented with gesture + speech combinations, but it is not clear if children and adults would benefit from multimodality to the same extent. Furthermore, it is

possible that adults can converge on the same interpretation in unimodal conditions (gesture-only, speech-only) as in the gesture + speech condition (e.g., when the interaction partner points at an open door without speech, adults may often correctly guess the intention and close the door), in which case their experimental paradigms are not suitable for comparison of multimodal gain in adults and children.

Another point that needs clarification is whether children integrate speech and gesture that mutually constrained each other's meanings. In Kelly (2001) and in Morford and Goldin-Meadow (1992), it is possible that all conditions conveyed the same message, but the message was vague and unclear in the unimodal conditions. In such cases, the benefit of multimodality could have been either the improvement of the signal-to-noise ratio or the opportunity to consult another modality when one modality is not clear. This contrasts with another type of multimodal benefit in which the messages in the unimodal condition are very clear, but when the two modalities are combined the meanings in the two modalities constrain each other to arrive at a unified message (as in the stimuli of the current study).

It is also not clear from the two studies at what age children can integrate into iconic gesture and speech. Kelly (2001) did not examine iconic gestures in his study. The study by Morford and Goldin-Meadow (1992) included two iconic gestures (throw and shake gesture), but the 14- to 28-month-old children did not understand iconic gestures without speech. When they were presented with a combination of two words (e.g., "Shake the book") or a combination of a gesture and a word(s) (e.g., "Book" + shake gesture or "Shake the book" + shake gesture), their comprehension of the gesture-speech combination was not better than their comprehension of the two-word sentence without the gesture (Morford & Goldin-Meadow, 1992).

Another reason why we focused on iconic gestures in this study is because we expected the results to be different from pointing gestures that indicate a physical object as a referent. This is because iconic gestures and pointing gestures are semiotically different, which leads to a processing load difference. A pointing gesture with a physical referent indicates its referent based on spatial contiguity to the referent. An iconic gesture represents its referent based on the resemblance to the referent. Pointing gestures with a physical referent can be interpreted, with help from rich information in the physical environment, whereas iconic gestures need to be interpreted based on features in the hand movement. Thus, processing iconic gesture may have a higher cognitive load than processing pointing gestures. In addition, previous studies on production of gestures have shown that pointing gestures emerged earlier than iconic gestures (e.g., Goldin-Meadow & Morford, 1990; Iverson, Capirci, & Caselli, 1994; Zinober & Martlew, 1985). Given the differences between iconic and pointing gestures, it is important to examine how and at what age people integrate two representations conveyed by iconic gesture and speech.

In summary, we aimed to investigate whether in children and/or adults iconic gesture and speech *mutually constrain* each other's meanings to form a unified message. We compared 3-year-olds, 5-year-olds, and adults to see whether 3-year-olds, whose understanding of gestures has not reached the adult-level, can make use of information from an iconic gesture to integrate it with information from speech.

The current study adapted the methods used with adults (Cocks, Morgan, & Kita, 2011; Cocks et al., 2009). In their studies, participants watched video vignettes of people producing short sentences in three different conditions, verbal only (V), gesture only (G) and verbal gesture combined (VG) (see Figs. 1, 2, and supporting information). The participants were required to select a picture that represented the intended meaning from one of four alternatives: integration match, a verbal-only match, a gesture-only match, and an unrelated foil. The integration match was the only correct response in the VG condition. The information in speech and the information gesture needed to constrain each other to arrive at the correct response in the VG condition. An index called "multi-modal gain" (MMG) was used to measure such integration. MMG indicates the degree to which the probability of choosing the integration match in VG goes beyond what is expected from the probabilities of choosing the integration match in V and G. MMG is defined in such a way that we can compare the degrees of integration by adults and young children who are expected to have different levels of performance in G (comprehension of gestures alone).

2. Experiment 1

2.2. Method

2.2.1. Participants

Twenty-four 3-year-olds (M_{age}: 3;8, range: 3;3 to 4;1, 12 females and 12 males), 24 5-year-olds (M_{age}: 5;7, range: 5;2 to 6;2, 12 females and 12 males), and 18 adults (M_{age}: 29, range: 21 to 38, 10 females and 8 males), who were all monolingual speakers of Japanese, participated.

2.2.2. Material

An actor was filmed producing eight combinations of an iconic gesture and a short sentence. The lower part of the actor's face was covered by a mask to conceal lip and mouth movement (see Fig. 1). This was done because the gestures in the videos in the

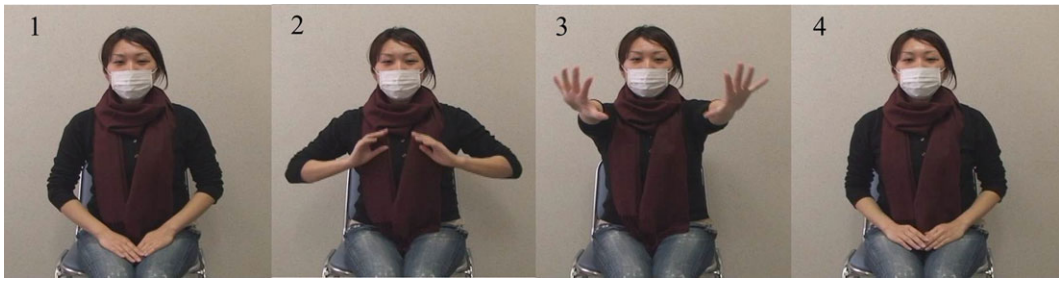

Fig. 1. An example of a stimulus gesture for "throwing."

Fig. 2. The four photographs as response choices shown after the "throwing" gesture in Fig. 1 was presented (upper left, verbal-only match; upper right, gesture-only match; bottom right, unrelated foil; bottom left, integration match).

VG condition and G condition were taken from recordings in which the actor with the mask both spoke and gestured, and without the mask, the lip movement would have given participants information about speech in the G condition. In the final stimuli, the speech produced by the actor with a mask was not used because the speech was somewhat muffled. The audio part of the final stimuli was separately recorded by the actor. It is highly common in Japan to wear a mask of the type used in the stimuli to avoid catching a cold or hay fever, and thus children were familiar with people wearing a mask of that type. The gestures expressed eight common everyday actions (writing, throwing,

riding, reading, drinking, eating, opening, and climbing; see Supporting Information), and the sentences referred to the same action (e.g., *nage-teimasu*, "(One) is throwing") with corresponding verbs (*kaku*, *nageru*, *noru*, *yomu*, *nomu*, *taberu*, *akeru*, and *noboru*, respectively). The eight verbs were selected based on the Japanese MacArthur Communicative Development Inventory (Watamaki & Ogura, 2004), and the verbs were understood by more than 80% of 3-year-olds. In the speech part of the stimulus, a subject noun phrase was dropped in the sentence, because a third-person pronoun is rarely used when talking to young Japanese children (it is grammatical in Japanese to omit the subject noun phrase; Shibatani, 1990). Gestures depicted one of possible ways (e.g., throwing a basketball) in which the action denoted by the verb (e.g., throwing) can be carried out (see Fig. 1 and the Supporting Information). From the recording of eight speech-gesture combinations, three versions of stimulus videos (totaling 24) were created for each speech-gesture combination: verbal + gesture (VG) (created by video editing to combine the audio part used in the V condition and the visual part used in the G condition), gesture-only (G) (the auditory part is muted), verbal-only (V) (the visual part is a still picture of the actor). Each video lasted 5 s.

Each speech-gesture combination had four corresponding color photographs as response choices (see Figs. 1, 2): (1) integration match, (2) verbal-only match, (3) gesture-only match, (4) and unrelated foil. Both integration matches and the verbal-only matches were semantically congruent with the speech and therefore were both correct choices in the V condition (but gesture-only matches and unrelated foils were not). In the G condition, both integration matches and gesture-only matches were semantically congruent with the gesture and therefore were correct choices (but the verbal-only match and the unrelated foil were not). In the VG condition if participants integrated the information from the speech and the gesture, the integration match was the only correct choice. Thus, the chance level for the V condition and the G condition is 50% each, but that for the VG condition is 25%. The four photographs for a given speech-gesture combination were arranged in a PowerPoint slide as in Fig. 2.

In the case of action "throw," in the V condition, the verbal stimulus was "(one) is throwing," and the correct choices were the picture of the boy throwing the basketball (integration match) and the picture of the boy throwing the baseball (verbal-only match). In the G condition, the gesture stimulus was a video clip showing an action of throwing by both hands without sound (Fig. 1), and the correct choice was the picture of the boy throwing the basketball (integration match) and the picture of the boy opening a door (gesture-only match). In the VG condition, the verbal stimulus was "(one) is throwing" and the gesture stimulus was a clip of throwing action, and the correct choice was only the picture of the boy throwing the basketball (integration match).

2.2.3. Stimulus presentation

Two of the eight actions (writing and throwing) in the three conditions (total six trials) were used in the practice trials. The participants watched six speech-gesture combinations in all three conditions, resulting in a total of 24 trials, in one of six different orders for counterbalancing. The order of six actions was consistent (riding, reading, drinking,

eating, opening, and climbing), and the three conditions were intermixed within the order of presentation (i.e., they were not blocked). The positions in which the trials for the three conditions appeared were counterbalanced across participants. The locations of the four types of response photographs were also counterbalanced so that each type of photograph appeared in each of the four locations on a slide equally often. In other words, the same set of four photographs was used for each action but the layout of the four photographs on the slide differed for each condition and the same four photographs reappeared with five intervening trials with the other sets of photographs.

2.2.4. Procedure

Participants were tested individually in a quiet spare room at the school for children and in a laboratory at the university for adults. At the beginning of the experiment, the experimenter instructed the participants as follows: "In this game you are going to see and hear a woman on this screen talking about different things people are doing. After you have watched the lady, you will see four different pictures. You need to pick one of the pictures that match what the lady was telling you about. Listen and watch carefully. Are you ready?" The Japanese original instruction can be seen in Appendix B. Participants were asked to watch a video stimulus embedded in a PowerPoint presentation on a laptop with a 12-inch screen, After each video stimulus, they saw four color photographs that simultaneously appeared on the screen for a response and were then asked by the experimenter to point to a photograph that best matched the message portrayed in the video as follows: "Can you point to one of the pictures that the lady was telling you about?" The same prompt was used in the three conditions. In an example of "throwing," a child is shown a video clip where the actor says "throwing" and gestures "throw basketball." The experimenter recorded every response on paper. Two practice trials were followed by experimental trials. In the two practice trials, if a participant pointed to the correct photograph, (s)he received positive feedback like "That's right" by the experimenter, whereas if (s)he pointed to the incorrect one, (s)he was given further prompts like "Oh do you think it is that one? What is he doing in that picture? Maybe it's a different picture. Which one do you think it might be?" This was repeated until they pointed to the correct photograph. Before starting the real trials, the experimenter confirmed whether the child understood the procedure. In the experimental trials, the child was always given positive feedback, regardless of his/her choice of the photograph. The experiment lasted, on average, 15 min.

2.2.5. Data analysis

We calculated an index of participants' ability to integrate verbal and gestural information (Cocks et al., 2009, 2011). In the VG condition, the integration match was the only correct response if the unique information in the verbal and gestural modalities were semantically integrated. However, the integration match could have been chosen even if the participant focused on one modality and ignored the other because the integration match was compatible with both speech and gesture. Therefore, successful choice of integration matches in the VG condition itself was not a good index of speech–gesture

integration. Thus, we needed to evaluate whether the probability of choosing integration matches the multimodal (VG) condition was higher than what would be expected from the probabilities of choosing integration matches in the unimodal (V, G) conditions. In other words, we need to know whether, in the VG condition, participants picked the integration match choice by putting information from both modalities into a unified interpretation or by making a decision based on only one modality. To this end, we calculated an index, MMG, for each participant.[1]

First, the proportions of integration match choice in the VG, V, and G conditions were calculated. The results from the V and G conditions indicated how often participants picked the integration match choice when only speech or only gesture was presented. Next, we estimated the upper bound of the probability at which participants would choose the integration match on the basis of a single modality (speech or gesture) in the VG condition. This was estimated as the maximum of the proportions of integration match choice in the V and G (i.e., whichever proportion that was bigger) for each participant. If participants used only a single modality in the VG condition, their performance could not exceed this maximum. Finally, the multimodal gain was estimated for each participant as the difference between the proportion of integration match choice in the VG condition and the maximum proportion in the V and G conditions, as in (1).

$$\begin{aligned}\text{MMG} = &\text{Prop. of integration match choice in VG} - \text{Maximum (Prop. of} \\ &\text{integration match choice in V, Prop. of integration match choice in G).}\end{aligned} \tag{1}$$

For example, out of six trials for each condition, a child picked four integration match choices for the V condition, two integration match choices for the G condition, and five integration match choices for the VG condition. The proportions of integration match choice in each condition were 0.67 (4/6), 0.33 (2/6), and 0.83 (5/6), respectively. Because the proportion of integration match choice in the V condition was higher than that in the G condition, the proportion in the V condition was used for calculation for the MMG score. Therefore, the child's MMG score is 0.16, by subtracting 0.67 from 0.83.

MMG represented an integration match choice in the VG condition that went beyond the estimated "best-case-scenario" performance if the participants used only one of the two modalities in the VG condition. The performance in the "best-case-scenario" was estimated to be the highest probability of integration match choice between the two unimodal conditions. Thus, MMG controls for participants' unimodal ability by taking a difference between the performance in the VG condition and the best performance in the V or G condition. A result where MMG = 0 would indicate that the participants could have picked the integration match by focusing on one modality. In other words, it indicates that multimodality did not contribute to the likelihood of the integration match choice in the VG condition. Conversely, a positive MMG score means that the choice of integration matches increased in the VG condition because participants integrated information from speech and gesture. Essentially the same variable is used in fMRI studies with multimodal and unimodal conditions (e.g., Straube, Green, Bromberger, & Tilo, 2011) to

isolate voxels representing brain areas for multimodal integration (e.g., (Speech + Gesture > Speech) ∩ (Speech + Gesture > Gesture); that is, voxels in which the activation in the multimodal condition is higher than in both unimodal conditions).[2] A similar measure of multimodal enhancement in performance in a behavioral task has also been proposed by Rach, Diederich, and Colonius (2011).

2.3. Results

No statistically significant differences were found in the proportion of trials with correct responses that children gave to the six presentation orders (3-year-olds, the V condition, $F(5, 18) = 1.16$, p = n.s., the G condition, $F(5, 18) = 0.94$, p = n.s., the VG condition, $F(5, 18) = 0.70$, p = n.s.; 5-year-olds, the V-condition, $F(5, 18) = 0.60$, p = n.s., the G condition, $F(5, 18) = 1.71$, p = n.s., the VG condition $F(5, 18) = 0.73$, p = n.s.; Adults, the V condition, $F(5, 12) = 1.00$, p = n.s., the VG condition, $F(5, 12) = 0.47$, p = n.s.). Data were therefore collapsed across presentation orders.

2.3.1. Correct choices in the three conditions

We examined whether the proportion of correct choices was above chance level, 50% for the V and G conditions and 25% for the VG condition. The proportions of trials with a correct choice were significantly higher than chance level for each condition for all age groups, one sample t-tests, ps < .05.

We conducted analyses of variance (ANOVA) for each of the three (V, G, VG) conditions on the mean proportion of trials with a correct choice with the age groups as a between-subject factor (Table 1). A main effect of age group was found for the G condition, $F(2, 63) = 21.44$, $p < .001$, $\eta^2 = .41$, and for the VG condition, $F(2, 63) = 43.08$, $p < .001$, $\eta^2 = .58$, but not for the V condition, $F(2, 63) = 2.36$, p = n.s. Fisher LSD post hoc tests (as suggested by Howell (2007) for comparison of the three means) showed that the proportions of trials with a correct choice were significantly higher in 5-year-olds and adults than in 3-year-olds in both G and VG conditions. This indicated that it is relatively difficult for 3-year-olds to comprehend information only from gesture, and from both speech and gesture.

2.3.2. Mean score of multimodal gain

To assess to what extent the participants semantically integrated information from speech and gesture in the VG condition, MMG scores were calculated for each partici-

Table 1
Mean proportion of correct choices chosen in each of the three conditions (V = verbal only, G = gesture only, VG = verbal gesture combined) and the standard deviations in parentheses

Condition	3 years	5 years	Adults	3 years in Experiment 2
V	0.95 (0.08)	0.98 (0.06)	0.99 (0.04)	0.94 (0.08)
G	0.74 (0.19)	0.92 (0.11)	1 (0.00)	0.73 (0.13)
VG	0.51 (0.22)	0.87 (0.14)	0.95 (0.10)	0.65 (021)

pant. An ANOVA was conducted on the MMG scores with age group as a between-subject factor (Fig. 3). A main effect of age group was found, $F(2, 63) = 3.19$. $p < .01$, $\eta^2 = .09$. Fisher LSD post hoc tests showed that 3-year-olds' MMG score was significantly lower than 5-year-olds ($p < .05$), suggesting that it is relatively difficult for 3-year-olds to semantically integrate information from gesture and speech. Note that 3-year-olds' poorer gesture comprehension in the G condition cannot account for their lower MMG scores as MMG indicates the degree to which people's performance in the VG condition goes beyond their performance in the G condition (and the V condition). That is, a low score in the G condition does not necessarily lead to a low MMG score.

A positive MMG score means that the choice of integration matches increased in the VG condition from that in the V or G condition, because participants integrated information from speech and gesture. MMG was significantly larger than zero in 5-year-olds, $t(23) = 3.46$, $p < .01$, $d = 0.71$, and adults, $t(17) = 2.95$, $p < .01$, $d = 0.70$, but not in 3-year-olds, $t(23) = .14$, p = n.s. Thus, only 3-year-olds did not obtain a significant increase in MMG scores in the VG condition from zero.

We also conducted an item-by-item analysis on MMG scores. In six out of six items, MMG scores in 5-year-olds were higher than those in 3-year-olds. In five out of six items, MMG scores in adults were equal to or higher than those in 3-year-olds. Thus, an item-by-item analysis showed the same tendency as the analysis of the MMG score aggregated over the six items.

2.3.3. Analysis of the failure of integration in 3-year-olds

It was revealed that the MMG score in 3-year-olds was lower than the other two age groups. This indicates that their ability to integrate speech and gesture was poorer than other age groups, above and beyond their poorer ability to pick up information from ges-

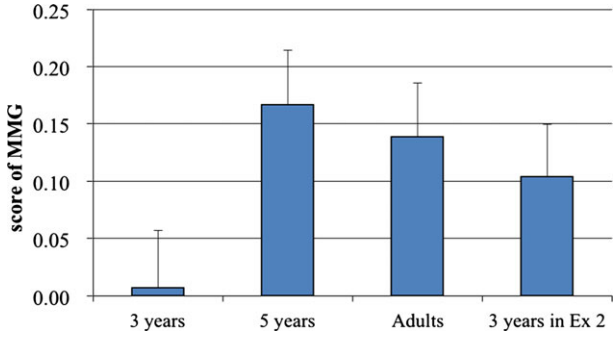

Fig. 3. Mean score of multimodal gain (MMG) in each age group. MMG is an index of how much multimodality contributed to the choice of integration matches in the VG (verbal + gesture) condition. MMG = 0 would indicate that the likelihood of choosing integration matches in the VG condition can be fully accounted for by the likelihoods of choosing integration matches in unimodal (verbal-only, gestural-only) conditions. Error bars represent standard errors of the means.

tures themselves. To further substantiate this point, we conducted an analysis that focused on items for which participants were successful in the G condition.

For each item (target action), we first selected participants who picked the correct choice in the G condition (either gesture-only match or integration match) (see Table 2). For this subset of participants, failure in the VG condition cannot be attributed to the inability to understand the gesture. Then, out of these participants, we calculated the proportion of participants (for each age group, for each target action) who picked the integration match in the VG condition (Table 3). To investigate whether these proportions differed between age groups, a chi-squared test was conducted for each target action. Significant differences were found for all target actions but *Reading*: *Riding*, $\chi^2(2, N = 57) = 14.48$, $p < .001$, *Reading*, $\chi^2(2, N = 61) = 3.78$, $p = .15$, *Drinking*, $\chi^2(2, N = 57) = 17.63$, $p < .001$, *Opening*, $\chi^2(2, N = 55) = 15.67$, $p < .001$, *Eating*, $\chi^2(2, N = 58) = 8.39$, $p < .001$, and *Climbing*, $\chi^2(2, N = 57) = 21.90$, $p < .001$. Post hoc analyses (with chi-squared tests comparing only two groups) indicated that for all target actions, the proportion of 3-year-olds who picked the integration match in the VG condition was significantly lower than the other two age groups. Thus, even when we focused on the cases where participants succeeded in the G condition, 3-year-olds performed worse in the VG condition than the other age groups. This result confirmed our claim that

Table 2
Mean proportion of participants who picked the correct choice in the G condition (within each age group) and the absolute number of participants in parentheses

	3 years	5 years	Adults
Riding	0.63 (15)	1.0 (24)	1.0 (18)
Reading	0.83 (20)	0.96 (23)	1.0 (18)
Drinking	0.83 (20)	0.83 (20)	0.94 (17)
Opening	0.63 (15)	0.92 (22)	1.0 (18)
Eating	0.75 (18)	0.92 (22)	1.0 (18)
Climbing	0.71 (17)	0.92 (22)	1.0 (18)

Table 3
Item analysis of the VG (verbal gesture combined) condition, focusing on cases in which participants correctly responded to the G (gesture-only) condition. Out of the participants who picked the correct choice (gesture match or integration match) in the G condition, the proportion of participants who picked the integration match in the VG condition (absolute number of participant is in parentheses) for each target action was calculated

	3 years	5 years	Adults
Riding	0.40 (6)	0.83 (20)	0.94 (17)
Reading	0.75 (15)	0.91 (21)	0.94 (17)
Drinking	0.35 (7)	0.70 (14)	1.0 (17)
Opening	0.40 (6)	0.77 (17)	1.0 (18)
Eating	0.66 (12)	0.95 (21)	0.94 (17)
Climbing	0.53 (9)	1.0 (22)	1.0 (18)

3-year-olds' poor integration of speech and gesture is not simply because they have difficulty in understanding gestures.

2.3.4. Error analysis

We analyzed the types of errors in the VG condition to see if verbal or gestural information dominated when the integration match was not chosen. We calculated the mean proportion of each type of photograph chosen in the VG condition for each age group (Table 4). We also calculated the mean proportion of each type of photograph chosen in the V condition and G condition, respectively, which are shown in Appendix C.

The participants rarely selected the unrelated foil. In addition, most adults did not make any errors. Thus, we compared the proportions of trials with verbal-only match choice with those with gesture-only match choice in the VG condition for 3- and 5-year-olds. Verbal-only matches were chosen significantly more often than gesture-only matches for 3-year-olds, $t(23) = 7.83$, $p < .001$, $d = 2.33$, and for 5-year-olds, $t(23) = 4.73$, $p < .001$, $d = 1.22$. This indicates that children tended to put more weight on verbal information than gestural information in the VG condition.

2.4. Discussion

Experiment 1 showed that, as we expected, when participants were shown only an iconic gesture, even 3-year-olds selected correct choices above the chance level, but their performance was lower than that of 5-year-olds and adults. It also indicated that when participants were shown both iconic gesture and speech, 3-year-olds did not integrate the unique information in the two modalities to arrive at a unified interpretation, and this difficulty cannot be attributed to their poorer ability to comprehend gestures. However, 5-year-olds showed an adult-like integration ability. Thus, for the first time, our study showed that iconic gesture and speech *mutually constrain* each other's meanings to arrive at a unified message in 5-year-olds and adults, but not in 3-year-olds. When 3-year-olds failed to integrate iconic gestures and speech, they tended to rely on information from speech.

One problem in interpreting the results of Study 1 is that it is unclear whether the 3-year-olds' performance on the VG stimuli was truly an integration failure rather than a consequence of the video presentation. We tested this problem by presenting the stimuli

Table 4
Mean proportion of each target chosen in the VG (verbal + gestural) condition for each age group, with the standard deviations in parentheses

Target Type	3 years	5 years	Adults	3 years Experiment 2
Verbal match	0.40 (0.20)	0.12 (0.12)	0.03 (0.06)	0.31 (0.20)
Gesture match	0.05 (0.09)	0.014 (0.05)	0.01 (0.04)	0.04 (0.10)
Integration target	0.51 (0.22)	0.87 (0.14)	0.95 (0.10)	0.65 (0.21)
Unrelated foil	0.02 (0.07)	0 (0)	0 (0)	0 (0)

in a different way from Experiment 1. In Experiment 2, like the live condition in Kelly's (2001) study, we used a naturalistic methodology in which an experimenter showed a live gesture and/or speech to 3-year-olds.

Children younger than 3-year-olds learn more from real-life experiences than from television. This phenomenon is termed the "video deficit effect" (Anderson & Pempek, 2005). This effect is observed in various tasks such as the imitation of an actor's action, object retrieval, word learning, and emotional response (e.g., Anderson & Pempek, 2005; Hayne, Herbert, & Simcock, 2003; McGuigan, Whiten, Flynn, & Horner, 2007; Richert, Robb, & Smith, 2011). In Experiment 2, we test whether 3-year-olds perform better if they are presented with the stimuli live by an experimenter.

In the context of the comprehension of gesture and speech, none of the previous studies have directly compared the effect of video versus live presentation, except for Kelly's (2001) study. Kelly showed young children's understanding of a pointing gesture and speech was better when they saw a speaker producing them live compared to when they saw her producing them in a video clip, and speculated that this is because young children are more sensitive to nonverbal behaviors when they are actual participants rather than observers of communicative interaction. Given this finding, observing a speaker in a real situation may have a different effect on the integration of speech and iconic gesture than from observing her or him in a video clip. Thus, Experiment 2 examined whether even 3-year-olds may show better performance in the integration of iconic gestures and speech when they see them produced live by an experimenter than the 3-year-olds in Experiment 1.

3. Experiment 2

3.1. Method

3.1.1. Participants

The participants were twenty-four 3-year-olds (M_{age}: 3;7, range: 3;1 to 4;0, 12 females and 12 males), who were all monolingual speakers of Japanese and did not participate in Experiment 1.

3.1.2. Material and procedure

The material, stimulus presentation, and procedure in Experiment 2 were the same as Experiment 1 except that an experimenter, who did not wear a mask, demonstrated a gesture and/or speech, instead of an actor in a video clips. After sitting down in front of a child, the experimenter demonstrated a target spoken sentence and/or gesture to a child. Children were instructed to look at four color photographs as responses on a laptop screen after the experimenter produced a target spoken sentence and/or gesture, then to point to the photograph that matched the message conveyed by the experimenter. The laptop was placed on the right side of each participant. The experiment lasted 15 min, on average.

3.2. Results

As in Experiment 1, no statistically significant differences were found in the pattern of responses children gave to the six presentation orders in any of the three conditions (the V condition, $F(5, 18) = 2.23$, p = n.s., the G condition, $F(5, 18) = 1.31$, p = n.s., the VG condition, $F(5, 18) = 0.09$, p = n.s.). Data were, therefore, collapsed across presentation orders.

3.2.1. Correct choices in the three conditions

We examined whether the proportion of correct choices of 3-year-olds in Experiment 2 was above the chance level, 50% for the V and G conditions and 25% for the VG condition. The proportions of trials with a correct choice were significantly higher than the chance level for each condition, one sample t-tests, ps < .05.

To examine the differences in performance between the 3-year-olds in Experiment 1 and those in Experiment 2, we conducted t-tests for each of the three conditions (V, G, and VG) on the mean proportion of trials with a correct choice (Table 1). A significant difference was found only in the VG condition: The proportion of trials with a correct choice from the 3-year-olds in Experiment 2 was significantly higher than those in Experiment 1, $t(23) = 2.22$, $p < .05$, $d = 0.42$. This indicated that it was easier for 3-year-olds to pick the integration match when gesture and speech were produced in the live condition.

3.2.2. Mean score of multimodal gain

Next, we investigated the difference of the MMG score between the 3-year-olds in Experiment 1 and those in Experiment 2 (Fig. 3). MMG score did not significantly differ between the Experiments, $t(23) = 2.16$, n.s. However, the MMG score was significantly larger than zero in the 3-year-olds in Experiment 2, $t(23) = 2.28$, $p < .05$, $d = 0.43$. Thus, there is some evidence that the live condition makes it easier for 3-year-olds to semantically integrate unique information in gesture and speech.

3.2.3. Error analysis

We analyzed the types of errors made by the 3-year-olds in Experiment 2 in the VG condition to see if verbal or gestural information dominated when the integration match was not chosen (Table 4). Just like the 5-year-olds and adults in Experiment 1, none of the 3-year-olds in Experiment 2 selected the unrelated foil. In the VG condition, verbal-only matches were chosen significantly more often than gesture-only matches, $t(23) = 5.38$, $p < .001$, $d = 0.75$. This indicates that even when gesture and speech were presented live, children tended to put more weight on verbal information than gestural information.

4. General discussion

To examine whether the development of speech–gesture integration is a part of a larger developmental shift occurred between 3- and 5-year-olds (Ramscar & Gitcho, 2007), we

investigated whether in children and adults iconic gesture and speech mutually constrain each other's meanings to form a unified message. In Experiment 1, we compared different age groups to see whether they can make use of information from an iconic gesture to integrate it with information from speech. In Experiment 2, we examined the performance of 3-year-olds in a live presentation of speech-gesture stimuli to see whether live interaction allows children to integrate iconic gestures and speech.

There are five main findings. First, when participants were shown only iconic gestures, even 3-year-olds selected correct choices above the chance level, but their performance was lower than that of 5-year-olds and adults. Second, 3-year-olds did not integrate the unique information in the two modalities to arrive at a unified interpretation when shown both an iconic gesture and a speech in video, but the 5-year-olds showed adult-like integration ability. Third, 3-year-olds' poor performance in the integration of speech and iconic gesture cannot be accounted for by their difficulty in interpreting the gesture. Fourth, when the children failed to integrate iconic gesture and speech, they tended to rely on information from speech. Fifth, when iconic gesture and speech are presented live, even 3-year-olds could integrate them.

The results indicate two things about development of children's ability to integrate gestures and speech. First, the development of speech–gesture integration may be a part of a broader developmental shift suggested by Ramscar and Gitcho (2007): The ability to process two different pieces of information at the same time develops between 3- and 5-year-olds (e.g., Deák, 2000; Frye et al., 1995; Loewenstein & Gentner, 2005; Zelazo et al., 2003). Though iconic gesture and speech have a special affinity with each other in the perceiver's mind (Kelly et al., 2014), this did not exempt iconic gesture and speech from the broader developmental shift.

Second, 3-year-olds were able to integrate iconic gesture and speech when gesture and speech are presented live. The current study indicated that 3-year-olds integrate speech and iconic gestures (MMG significantly above zero) in the VG condition when speech and gesture are presented live (Experiment 2), but not when they are presented in a video (Experiment 1). This result is consistent with previous studies showing that young children performed less well from video presentations of stimuli than from equivalent real-life experiences (Anderson & Pempek, 2005; Hayne, Barr, & Herbert, 2003; Richert et al., 2011). Examining the reason why a live condition makes it easier for young children to integrate gesture and speech would be an important topic for future research.

The 3-year-olds' above-chance performance in the gesture-only condition is in line with previous findings that even infants can pick up information encoded only in iconic gestures (McGregor et al., 2009; Tomasello, Striano, & Rochat, 1999). However, 3-year-olds' performance was significantly worse than 5-year-olds and adults. This may be because our stimulus gestures required children to imagine an object associated with a gesturally depicted action. Previous studies have shown that 3-year-olds were worse at comprehending gestures with an imaginary object (e.g., a gesture for brushing teeth with a fist handshape holding an imaginary tooth brush) than 4- and 5-year-olds and adults (e.g., Boyatzis & Watson, 1993; O'Reilly, 1995; Overton & Jackson, 1973). This difficulty in gesture comprehension may be the reason why 3-year-olds rely on the

information conveyed in speech rather than gesture when they make an error in the VG condition. Consistent with this interpretation, Zelazo et al. (2003) argued that when 3-year-olds are simultaneously presented with two representations, they tend not to integrate the two representations and perseverate on a salient representation (in the current study, speech).

This study also makes a methodological contribution. Despite the age difference in the ability to comprehend iconic gestures, the MMG score is still a useful measurement for speech and gesture integration. This is because MMG measures how far participants went beyond their unimodal ability by taking the difference between the performance in the VG condition and the best performance in the V or G condition (see Straube et al., 2011; Rach et al., 2011 for similar measures). Another strength of MMG is that it is a general variable, applicable to all possible contexts of speech–gesture integration. We can distinguish three contexts in which speech–gesture integration can be observed: (a) when speech is not easily interpretable, (b) when gesture is not easily interpretable, and (c) when speech and gesture are perfectly interpretable. The second context is the situation that our 3-year-olds were in, and the third context is the situation that our 5-year-olds and adult participants were in. It would be interesting for future research to compare how 3-year-olds perform in all three contexts. This study is important because it opens doors for a further systematic investigation of development of speech–gesture integration in comprehension.

To conclude, the current study developed a novel method to investigate children's ability to semantically integrate information from speech and iconic gestures. The results indicate that development of this integration ability may be part of a broader developmental shift between 3- and 5-year-olds (Ramscar & Gitcho, 2007), with regard to the ability to process two different pieces of information at the same time.

Notes

1. The calculation of MMG score in the current study is simpler and makes fewer assumptions about the nature of the integration process as compared to the calculation proposed by Cocks et al. (2009).
2. Straube, Green, Bromberg, and Kircher (2011) used slightly more elaborate conjunctions (Speech + Gesture > Speech ∩ Speech + Gesture > Gesture ∩ Speech ∩ Gesture) to make sure that the "integration areas" were also activated by unimodal information.

References

Anderson, D. R., & Pempek, T. A. (2005). Television and very young children. *The American Behavioral Scientist*, *48*, 505–522.

Boyatzis, C. J., & Watson, M. W. (1993). Preschool children's symbolic representation of objects through gestures. *Child Development*, *64*, 729–735.

Broaders, S. C., & Goldin-Meadow, S. (2010). Truth is at hand: How gesture adds information during investigative interviews. *Psychological Science*, *21*(5), 623–628.

Cassell, J., McNeill, D., & McCullough, K.-E. (1999). Speech–gesture mismatches: Evidence for one underlying representation of linguistic and nonlinguistic information. *Pragmatics & Cognition*, *7*, 1–34.

Church, R. B., Kelly, S. D., & Lynch, K. (2000). Multi-modal processing over development: The case of speech and gesture detection. *Journal of Nonverbal Behavior*, *24*, 151–174.

Cocks, N., Morgan, G., & Kita, S. (2011). Iconic gesture and speech integration in younger and older adults. *Gesture*, *11*, 24–39.

Cocks, N., Sautin, L., Kita, S., Morgan, G., & Zlotowitz, S. (2009). Gesture and speech integration: An exploratory study of a man with aphasia. *International Journal of Language and Communication Disorders*, *44*, 795–804.

Deák, G. O. (2000). The growth of flexible problem solving: Preschool children use changing verbal cues to infer multiple word meanings. *Journal of Cognition and Development*, *1*, 157–191.

Derks, P., & Paclisanu, M. (1967). Simple strategies in binary prediction by children and adults. *Journal of Experimental Psychology*, *73*, 278–285.

Flavell, J. H., Green, F. L., & Flavell, E. R. (1986). Development of knowledge about the appearance-reality distinction. *Monographs of the Society for Research in Child Development*, *51*, Serial No. 212.

Frye, D., Zelazo, P. D., & Palfai, T. (1995). Theory of mind and rule-based reasoning. *Cognitive Development*, *10*, 483–527.

Goldin-Meadow, S., & Morford, M. (1990). Gesture in early child language. In V. Volterra & C. J. Erting (Eds.), *From gesture to language in hearing and deaf children* (pp. 249–262). New York: Springer-Verlag.

Goldin-Meadow, S., & Sandhofer, C. M. (1999). Gestures convey substantive information about a child's thoughts to ordinary listeners. *Developmental Science*, *2*(1), 67–74.

Goodrich, W., & Hudson-Kam, C. L. (2009). Co-speech gesture as input in verb learning. *Developmental Science*, *12*(1), 81–87.

Hayne, H., Barr, R., & Herbert, J. (2003). The effect of prior practice on memory reactivation and generalization. *Child Development*, *74*, 1615–1627.

Hayne, H., Herbert, H., & Simcock, G. (2003). Imitation from television by 24- and 30-month-olds. *Developmental Science*, *6*, 254–261.

Howell, D. C. (2007). *Statistical methods for psychology* (6th ed.). Belmont, CA: Duxbury.

Iverson, J. M., Capirci, O., & Caselli, M. C. (1994). From communication to language in two modalities. *Cognitive Development*, *9*, 23–43.

Kelly, S. D. (2001). Broadening the units of analysis in communication: Speech and nonverbal behaviours in pragmatic comprehension. *Journal of Child Language*, *28*, 325–349.

Kelly, S. D., & Church, R. B. (1998). A comparison between children's and adults' ability to detect children's representational gestures. *Child Development*, *69*, 85–93.

Kelly, S. D., Healey, M., Özyürek, A., & Holle, J. (2014). The processing of speech, gesture, and action during language comprehension. *Psychonomic Bulletin & Review*, Advance online publication. doi:10.3758/s13423-014-0681-7. Published online, 08 July.

Kelly, S. D., Kravitz, C., & Hopkins, M. (2003). Neural correlates of bimodal speech and gesture comprehension. *Brain and Language*, *89*, 253–260.

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.

Loewenstein, J., & Gentner, D. (2005). Relational language and the development of relational mapping. *Cognitive Psychology*, *50*, 315–353.

McGregor, K. K., Rohlfing, K. J., Bean, A., & Marschner, E. (2009). Gesture as a support for word learning: The case of under. *Journal of Child Language*, *36*, 807–828.

McGuigan, N., Whiten, A., Flynn, E., & Horner, V. (2007). Imitation of causally opaque versus causally transparent tool use by 3- and 5-year-old children. *Cognitive Development*, *22*, 353–364.

McNeil, N. M., Alibali, M. W., & Evans, J. L. (2001). The role of gesture in children's comprehension of spoken language: Now they need it, now they don't. *Journal of Nonverbal Behavior*, *24*(2), 131–150.

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

Morford, M., & Goldin-Meadow, S. (1992). Comprehension and production of gesture in combination with speech in one-word speakers. *Journal of Child Language*, *19*(3), 559–580.

Mumford, K. H., & Kita, S. (2014). Children use gesture to interpret novel verb meanings. *Child Development*, *85*, 1181–1189.

Namy, L. L., Campbell, A. L., & Tomasello, M. (2004). The changing role of iconicity in non-verbal symbol learning: A U-shaped trajectory in the acquisition of arbitrary gestures. *Journal of Cognition and Development*, *5*(1), 37–57.

O'Reilly, A. W. (1995). Using representations: Comprehension and production of actions with imagined objects. *Child Development*, *66*, 999–1010.

Overton, W. F., & Jackson, J. P. (1973). The representation of imagined objects in action sequences: A developmental study. *Child Development*, *44*, 309–314.

Rach, S., Diederich, A., & Colonius, H. (2011). On quantifying multisensory interaction effects in reaction time and detection rate. *Psychological Research*, *75*, 77–94.

Ramscar, M., & Gitcho, N. (2007). Developmental change and the nature of learning in childhood. *Trends in Cognitive Science*, *11*, 274–279.

Resches, M., & Pérez Pereira, M. (2007). Referential communication abilities and Theory of Mind development in preschool children. *Journal of Child Language*, *34*, 21–52.

Richert, R. A., Robb, M., & Smith, E. (2011). Media as social partners: The social nature of young children's learning from screen media. *Child Development*, *82*, 82–95.

Shibatani, M. (1990). *The languages of Japan*. Cambridge, England: Cambridge University Press.

Straube, B., Green, A., Bromberg, B., & Kircher, T. (2011). The differentiation of iconic and metaphoric gestures: Common and unique integration processes. *Human Brain Mapping*, *32*, 522–533.

Tomasello, M., Striano, T., & Rochat, P. (1999). Do young children use objects as symbols? *British Journal of Developmental Psychology*, *17*, 563–584.

Watamaki, T., & Ogura, T. (2004). *Technical manual of the Japanese MacArthur communicative development inventory: Words and grammar*. Kyoto: Kyoto International Social Welfare Exchange Center.

Zelazo, P. D., Muller, U., Frye, D., & Marcovitch, S. (2003). The development of executive function in early childhood. *Monographs of the Society for Research in Child Development*, *68*, Serial No. 274.

Zinober, B., & Martlew, M. (11985). Developmental changes in four types of gestures in relation to acts and vocalizations from 10 to 21 months. *British Journal of Developmental Psychology*, *3*, 293–306.

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Data S1: Gestures and speech in the stimulus video clips and the description of each response choices.

Appendix A: Summary of previous studies on comprehension of gestures

Table A1
Summary of previous studies on comprehension of gestures in children

Author(s) and publication year	Age groups of participants	Comparison of adults	Gesture type used	Way of presenting the stimuli	Stimulus modalities: Used all three conditions (speech-only, gesture-only, and speech–gesture combination)?	Integration of speech and gesture required: Do speech and gesture mutually constrain each other's meaning?	Results regarding speech–gesture integration
Morford & Goldin-Meadow (1992)	14–28 month	No	Iconic, Pointing	Live	Yes	Yes (However, this study cannot exclude the possiblity that the belter performance in the speech + gesture condition was obtained from the advantage of redundancy between the speech and gesture modalities)	Children could integrate pointing gestures and speech by an adult speaker to select the correct object, but could not integrate iconic gestures and speech
Namy, Campbell, and Tomasello (2004)	18, 26 month, and 4-year-olds	No	Iconic, Arbitrary	Live	No (speech–gesture combination)	No (Participants were required to associate a gesture with an object. Speech does not covey information about the object. Speech and gesture do not constrain each other's meaning)	

(continued)

Table 5. (continued)

Author(s) and publication year	Age groups of participants	Comparison of adults	Gesture type used	Way of presenting the stimuli	Stimulus modalities: Used all three conditions (speech-only, gesture-only, and speech–gesture combination)?	Integration of speech and gesture required: Do speech and gesture mutually constrain each other's meaning?	Results regarding speech–gesture integration
Tomasello, Striano, & Roc hat (1999)	18 month, 2- and 3-year-olds	No	Iconic	Live	No (speech–gesture combination)	No (Same as above)	
Kelly & Church (1998)	10-year-olds and adults	Yes	Iconic, Pointing	Video	No (speech-only, speech–gesture combination)	No (Participants were required to describe information conveyed by children cither with speech or gesture. Information in speech and in gesture are equivalent or non-equivalent. Speech and gesture do not constrain each other's meaning)	

(continued)

Table 5. (continued)

Author(s) and publication year	Age groups of participants	Comparison of adults	Gesture type used	Way of presenting the stimuli	Stimulus modalities: Used all three conditions (speech-only, gesture-only, and speech–gesture combination)?	Integration of speech and gesture required: Do speech and gesture mutually constrain each other's meaning?	Results regarding speech–gesture integration
Church et al. (2000)	8-, and 10-year-olds, and adults	Yes	Iconic, Pointing	Video	No (speech-only, speech–gesture combination)	No (Participants were required to recognize information conveyed by children either with speech or gesture. Information in speech and in gesture are equivalent or non-equivalent. Speech and gesture do not constrain each other's meaning)	
Broaders & Goldin-Meadow (2010)	6-year-olds	No	Iconic	Live	No (speech-only, speech–gesture combination)	No (Participants were required to answer the question, about a person who they had seen, by an interviwer producing speech and gesture. Information in speech and in gesture are equivalent or non-equivalent. Speech does not constrain gesture's meaning)	

(continued)

Table 5. (continued)

Author(s) and publication year	Age groups of participants	Comparison of adults	Gesture type used	Way of presenting the stimuli	Stimulus modalities: Used all three conditions (speech-only, gesture-only, and speech–gesture combination)?	Integration of speech and gesture required: Do speech and gesture mutually constrain each other's meaning?	Results regarding speech–gesture integration
McGregor et al., 2009	20–24 month	No	Iconic	Live	No (speech–gesture combination)	No (Participants were required to use gesture to infer the meaning of the word, "under." After training, information in speech and in gesture are equivalent. Speech does not constrain gesture)	
Goodrich & Hudson-Kam (2009)	2-, 3-, and 4-year-olds, and adults	Yes	Iconic, Interactive	Live	No (gesture-only, speech–gesture combination)	No (Participants were required to use information in gestures to assign the meaning to a novel verb. After training, information in speech and in gesture are equivalent. Speech docs not constrain gesture)	

(continued)

Table 5. (continued)

Author(s) and publication year	Age groups of participants	Comparison of adults	Gesture type used	Way of presenting the stimuli	Stimulus modalities: Used all three conditions (speech-only, gesture-only, and speech–gesture combination)?	Integration of speech and gesture required: Do speech and gesture mutually constrain each other's meaning?	Results regarding speech–gesture integration
McNeil, Alibali, & Evans (2001)	4- and 5-year-olds, 5- and 6-year-olds	No	Iconic, Pointing	Video	No (speech-only, speech–gesture combination)	No (Participants were required to select a block according to the instructions given by a speaker with gesture and speech. Information in speech and gesture are either equivalent or non-equivalent. Gesture affects speech in the non-equivalent condition, but speech docs not constrain gesture)	
Mumford & Kita (2014)	3-year-olds	No	Iconic	Live	No (speech-only, speech–gesture combination)	No (Participants were required to use gesture to infer the meaning of the word, "under." After training, information in speech and in gesture are equivalent. Speech does not constrain gesture)	

(continued)

Table 5. (continued)

Author(s) and publication year	Age groups of participants	Comparison of adults	Gesture type used	Way of presenting the stimuli	Stimulus modalities: Used all three conditions (speech-only, gesture-only, and speech–gesture combination)?	Integration of speech and gesture required: Do speech and gesture mutually constrain each other's meaning?	Results regarding speech–gesture integration
Kelly (2001)	3- and 4-year-olds, 4- and 5-year-olds	No	Pointing	Video (Ex 1), Live (Ex 2)	Yes	Yes (However, this study cannot exclude the possiblity that the better performance in the speech + gesture condition was obtained from the advantage of redundancy between the speech and gesture modalities)	Children could integrate pointing gestures and speech by an adults speaker to understand an actor's or the speakers indirect request
Current study	3-, 5-year-olds, and adults	Yes	Iconic	Video (Ex 1), Live (Ex 2)	Yes	Yes	Children could integrate iconic gestures and speech by an adult speaker to select the picture depicting the correct action

Appendix B: Original Japanese instruction.

このゲームでは、この画面にお姉さんが出てきて，ある人のお話をします。お姉さんがお話をした後、4つの写真がでてきます。そしたら，お姉さんがお話をしていた人を、4つの写真のなかから選んでください。それでは、お姉さんのことをよくみて，お姉さんのお話をよく聞いてくださいね。準備はいいですか？

Appendix C: Mean proportion of each target (photograph) type.

Table C1
Mean proportion of each target chosen in the V (verbal) condition for each age group and the standard deviations in parentheses

Target type	3 years	5 years	Adults	3 years in Experiment 2
Verbal match	0.48 (0.16)	0.40 (0.23)	0.32 (0.17)	0.40 (0.22)
Gesture match	0.02 (0.07)	0.02 (0.06)	0.01 (0.04)	0.06 (0.09)
Integration match	0.47 (0.14)	0.58 (0.25)	0.67 (0.19)	0.53 (0.23)
Unrelated foil	0.03 (0.07)	0.00 (0.00)	0.00 (0.00)	0.02 (0.06)

Table C2
Mean proportion of each target chosen in the G (gestural) condition for each age group and the standard deviations in parentheses

Target type	3 years	5 years	Adults	3 years in Experiment 2
Verbal match	0.17 (0.16)	0.07 (0.10)	0.01 (0.04)	0.19 (0.14)
Gesture match	0.38 (0.19)	0.40 (0.26)	0.24 (0.17)	0.44 (0.19)
Integration match	0.36 (0.17)	0.53 (0.28)	0.76 (0.17)	0.28 (0.19)
Unrelated foil	0.10 (0.13)	0.01 (0.03)	0.00 (0.00)	0.06 (0.10)