COGNITIVE SCIENCE
A Multidisciplinary Journal

Cognitive Science 44 (2020) e12868

ISSN: 1551-6709 online
DOI: 10.1111/cogs.12868

Making Sense of the Hands and Mouth: The Role of "Secondary" Cues to Meaning in British Sign Language and English

Pamela Perniss,[a] David Vinson,[b] Gabriella Vigliocco[b]

[a]*Faculty of Human Sciences, University of Cologne*
[b]*Division of Psychology and Language Sciences, University College London*

Received 18 September 2018; received in revised form 1 May 2020; accepted 6 May 2020

Abstract

Successful face-to-face communication involves multiple channels, notably hand gestures in addition to speech for spoken language, and mouth patterns in addition to manual signs for sign language. In four experiments, we assess the extent to which comprehenders of British Sign Language (BSL) and English rely, respectively, on cues from the hands and the mouth in accessing meaning. We created congruent and incongruent combinations of BSL manual signs and mouthings and English speech and gesture by video manipulation and asked participants to carry out a picture-matching task. When participants were instructed to pay attention only to the primary channel, incongruent "secondary" cues still affected performance, showing that these are reliably used for comprehension. When both cues were relevant, the languages diverged: Hand gestures continued to be used in English, but mouth movements did not in BSL. Moreover, non-fluent speakers and signers varied in the use of these cues: Gestures were found to be more important for non-native than native speakers; mouth movements were found to be less important for non-fluent signers. We discuss the results in terms of the information provided by different communicative channels, which combine to provide meaningful information.

Keywords: Sign language; Mouthings; Audio-visual speech; Gesture; Integration; Primary and secondary cues; Multimodality; Language learners

Correspondence should be sent to Pamela Perniss, Faculty of Human Sciences, University of Cologne, Klosterstr. 79b, 50931 Cologne, Germany. E-mail: pperniss@uni-koeln.de

1. Introduction

Introductions to signed and spoken languages typically mention the radical difference in production and perception between the two language modalities, assigning a main and different articulatory organ in each case: Spoken languages are produced by the vocal tract and perceived by ear; signed languages are produced manually and perceived by eye. Yet, when we look at face-to-face interaction, it is clear that both modalities involve a range of different articulators and channels of expression. The speech signal is invariably accompanied by visual cues from the mouth, face, hands, and body. Similarly, manual signs are produced together with mouth, face, and body movements that contribute to utterance meaning.

The information conveyed in different channels and by different articulators exhibits tight semantic and temporal alignment. Iconic co-speech gestures, for example, occur in close alignment with their lexical affiliate (e.g., a throwing gesture occurring with the word *throw*; Church, Kelly, & Holcombe, 2013; Kendon, 1972; Loehr, 2007) and beat gestures co-occur with focused or prominent information in speech (e.g., Dimitrova, Chu, Wang, Özyürek, & Hagoort, 2016; Krahmer & Swerts, 2007; Leonard & Cummins, 2011). Similarly, facial movements, for example, brow movements, are closely coordinated with speech, especially with prosodic cues marking focus and prominence (Krahmer & Swerts, 2004), and the visible movements of the mouth are necessarily time locked with the phonetic articulation of speech. In signed language, mouth actions, including *mouthings* derived from the surrounding spoken language (e.g., silent articulation of the English word "apple" while producing the British Sign Language (BSL) sign APPLE[1] manually) occur in semantic and temporal relationship with corresponding manual productions (Bank, Crasborn, & van Hout, 2011; Sutton-Spence & Day, 2001). Other cues on the face, for example, raised or furrowed brows, mark grammatical information related to sentence structure and type, including topicalization and question marking (Liddell, 1980; Sutton-Spence & Woll, 1999), with scope indicated by clearly timed onsets and offsets (Pyers & Emmorey, 2008).

In both language modalities, language comprehension in face-to-face contexts thus involves the integration of multiple types of meaningful information. For spoken language, evidence for the use of information from different channels comes from audio-visual speech (e.g., McGurk & MacDonald, 1976; van Wassenhove, Grant, & Poeppel, 2005) and co-speech gesture (e.g., Habets, Kita, Shao, Özyürek, & Hagoort, 2011; He et al., 2015; Holle & Gunter, 2007; Kelly, Healey, Özyürek, & Holler, 2014; Kelly, Özyürek, & Maris, 2010; Obermeier, Kelly, & Gunter, 2015; Özyürek, Willems, Kita, & Hagoort, 2007; Straube, Green, Bromberger, & Kircher, 2011), or explicitly from both cues studied in a joint context (Drijvers & Özyürek, 2017, 2018, 2019). This body of research offers support, both behaviorally and from neuroimaging, for the automatic, simultaneous and bidirectional integration of information from speech and gesture. For example, Kelly et al. (2010) show that speakers cannot help but pay attention to information in gesture, even when gesture is not relevant to the task at hand, and furthermore, that speech and gesture are bidirectionally integrated—that is, incongruent speech and

incongruent gesture are equally disruptive to comprehension. Using an ERP paradigm, Özyürek et al. (2007) provide evidence that speech and gesture are simultaneously integrated into a preceding sentence context: speech, gesture, or speech + gesture mismatching the semantic context elicited the same effect on the N400 component in terms of latency and amplitude.

Other research has focused less directly on the integration of channels and more on facilitation or enhancement effects in difficult conditions. Both information from co-speech gesture and from mouth movements play an important role here. Visual information from the mouth, known to contribute greatly to speech perception (McGurk & McDonald, 1976), can facilitate processing of the auditory signal (van Wassenhove et al., 2005) and enhance speech comprehension in conditions of degraded speech (Ross, Saint-Amour, Leavitt, Javitt, & Foxe, 2007). Similarly, information from co-speech gesture has been found to enhance speech comprehension under adverse listening conditions (Holle, Obleser, Rueschemeyer, & Gunter, 2010; Obermeier, Dolk, & Gunter, 2012). Recent work by Drijvers and Özyürek (2017, 2018, 2019) has studied the enhancement effect of both cues jointly—phonological support from visible mouth movements and semantic support from co-speech gesture. Using free or cued recall tasks, Drijvers and Özyürek show that both cues together, that is, gesture and visible mouth movements, offer the greatest benefit to speech comprehension in degraded conditions: a "double enhancement" effect when both cues are present. The enhancement effect is moreover greater for moderately degraded compared to severely degraded speech, suggesting that there needs to be a certain (sufficient) amount of semantic information available from speech in order for gesture to be useful. While the focus of this research is on the enhancement of degraded speech perception through concurrently available visible cues, the findings suggest that gestures are actively processed and integrated with the speech signal.

Another line of research that addresses the relationship between speech and gesture for comprehension has focused on populations of language learners. Gestures have been shown to facilitate non-native speaker comprehension and to help in foreign language learning (e.g., Drijvers & Özyürek, 2018, 2019; Kelly, McDevitt, & Esch, 2009; Macedonia, Müller, & Friederici, 2011; Sueyoshi & Hardison, 2005). For example, Sueyoshi and Hardison (2005) found that Japanese speakers with low proficiency in English showed better comprehension of lectures when the speaker's face and gestures were visible to them compared to lectures in which only the speaker's face was visible or in which only the audio was provided. Kelly et al. (2009) found that native speakers of English being taught Japanese verbs learned best when verbs were accompanied by iconic gestures depicting the meaning of the verb compared to when speakers learned the verb only. Drijvers and Özyürek (2018) found support for larger reliance on gestures by non-native compared to native speakers in an EEG experiment: Non-natives showed a larger N400 effect than natives in clear speech, suggesting an increased recruitment of the visual semantic information from gestures by non-native speakers. In learners as well, the benefit from semantic cues in gestures seems to depend on the availability of auditory cues from speech. Drijvers and Özyürek (2018) found that the benefit of additional information from gesture held for non-native speakers only when speech was not too severely

degraded. That is, information from gesture was used only when information from speech could be sufficiently processed. Finally, it is interesting to note that the support from gestures for non-native speakers seems to be primarily on the level of meaning, in providing additional semantic cues, rather than on the phonological level. Hirata, Kelly, Huang, and Manansala (2014) found no support from gestures for the perception of novel phonological contrasts in language learning: English learners of Japanese did not benefit from gestures (iconically) representing long and short vowel contrasts in Japanese syllables and morae.

Compared to spoken language, our knowledge of the interplay and integration of different cues, and about the role of different channels as potentially modulated by language proficiency, is very limited for signed language. In signing, the hands are considered to be the predominant channel: Most lexical and grammatical expression takes place through movement and placement of the hands. However, non-manual channels are ubiquitous accompaniments to manual productions. Of particular interest to the present study are mouthings, that is, mouth movements derived from the surrounding spoken language and visually resembling the articulation of words (Sutton-Spence, 1999; Sutton-Spence & Woll, 1999). Although some mouthings serve to disambiguate otherwise similar signs (e.g., the signs AUNT and BATTERY in BSL which differ only in mouthing), they are frequently produced for unambiguous signs (e.g., 69% of all signs in a BSL corpus, Sutton-Spence, 2007; Sutton-Spence & Day, 2001) and do not independently contribute meaning (e.g., TABLE with mouthing "table," as compared to instances like PULLOVER accompanied by mouthing "red," see Vogt-Svendsen, 2001). Mouthings co-occur with manual signs, in terms of being temporally and semantically aligned, across a range of different sign types—including nouns, adjectives, and simple verbs—though the consistency of their occurrence varies both across signers and sign languages (e.g., Bank et al., 2011; Crasborn, van der Kooij, Waters, Woll, & Mesch, 2008; Johnston, van Roekel, & Schembri, 2016; Nadolske & Rosenstock, 2007). Spreading of mouthings over adjacent signs, especially pointing signs (Bank, Crasborn, & van Hout, 2013; Crasborn et al., 2008), is also common and has been argued to mark prosodic domains (Sandler, 1999). A central question with respect to mouthings has been whether they are an integral part of the phonological and lexical representation of signs or whether they constitute a separate phonological representation, to be analyzed as simultaneous code blending (see the papers in Boyes-Braem & Sutton-Spence, 2001 for arguments from both sides of the debate).[2] For the most part, studies of both production and comprehension have provided support for the position that mouthings reflect knowledge of two languages (Ebbinghaus & Heßmann, 2001; Giuostolisi, Mereghetti, & Cecchetto, 2017; Vinson, Thompson, Skinner, Fox, & Vigliocco, 2010; for evidence from neuroimaging see Capek et al., 2009). However, the behavior of mouthings on these different levels—prosodic, grammatical, and lexical—suggests that they are integrally constitutive of sign language use, and the result of complex processes of cross-modal blending and language contact (Bank, Crasborn, & van Hout, 2016; Mohr, 2012; van de Sande & Crasborn, 2009).

Despite the ubiquity of mouthings, no studies have examined how they are integrated with other components of signed language in comprehension. As with spoken language,

eye gaze is primarily on the face in sign language comprehension (Emmorey, Thompson, & Colvin, 2009; Muir & Richardson, 2005). Mouthings are thus readily accessible to sign comprehenders and support meaning expression, providing visually salient cues that help disambiguate or clarify the signs being produced manually (Bank et al., 2011). In line with this, the information available from mouthings has been shown to facilitate comprehension between signers who use different regional variants of a sign language (e.g., different variants of BSL, Stamp, 2016; Stamp, Schembri, Evans, & Cormier, 2016) as well as between signers who use different sign languages (e.g., Flemish Sign Language [VGT] and Sign Language of the Netherlands [NGT], which use mouthing derived from the same spoken language, Sáfár et al., 2015). In these situations, signers rely on mouthing for comprehension and to enhance mutual understanding. The regular co-occurrence of mouthings with manual signs in natural signing means that these cues are highly correlated. However, beyond the regional variation and mutual intelligibility studies mentioned above, we know little about how mouthing affects comprehension. Does this highly accessible visual cue support comprehension in contexts of shared manual signs? Moreover, does mouthing differentially influence comprehension for different types of signers: native deaf signers, native hearing signers, and learners of a sign language? Mouthings may play a different role for deaf versus hearing native signers due to being based on an auditory phonological representation or not. For hearing sign language learners, mouthings may play a greater role as a cue familiar from their L1. That is, the use of mouthing derived from English may be helpful to learners of BSL in providing a mouthed version of the English translation equivalent with the manual sign. On the other hand, because mouthings are seldom explicitly taught in formal sign language instruction and speechreading abilities in hearing non-signing adults are relatively poor (Mohammed, Campbell, Macsweeney, Barry, & Coleman, 2006; Pimperton, Ralph-Lewis, & MacSweeney, 2017), learners may use them less effectively. Learners of BSL may instead be focused on accessing meaning from the hands—the primary signal—and on learning the manual form of the signs, and may ignore the additional information on the mouth.

In the present paper, we investigate the nature of interaction between the hands and mouth as important articulators in both modalities. The study is novel in providing a systematic and comprehensive investigation of how a primary (speech; manual component of sign) and a secondary (gesture; mouthing) channel are integrated, and whether language proficiency modulates the use of secondary cues in comprehension. It is the first study to look at how the two channels are integrated in a signed language, BSL, and provides a replication and extension of previous work on speech and gesture.

We follow a paradigm similar to Kelly et al. (2010), constructing incongruent pairings of manual signs and mouthings (for BSL) and of audio-visual speech and gesture (for English) from congruent productions. We use the same method of stimulus material creation for both modalities, with a video editing method (see Section 2.1.2) that allowed us to construct more ecologically valid stimuli in which the face is visible and mouth movements correspond to heard speech. Kelly et al. (2010) and other earlier speech–gesture studies that constructed incongruent stimuli from semantically congruent productions used just the audio signal of speech (e.g., Habets et al., 2011; Holle & Gunter, 2007; Kelly

et al., 2014; Obermeier et al., 2012; Özyürek et al., 2007; Willems, Özyürek, & Hagoort, 2007; Wu & Coulson, 2014). The face was obscured to avoid an audio track mismatching the visual information available from lips during articulation. Drijvers and Özyürek (2018) use an incongruence paradigm with the face visible, similarly to the present study. However, in their study, incongruent speech–gesture pairings were obtained by asking the video model to produce words with mismatching gestures, thus creating conditions for less natural production.

We use a picture–video matching task in which participants see a picture followed by a video of an actor producing a word/sign accompanied by an iconic gesture/mouthing that is either congruent or incongruent with the word/sign. We test whether the secondary channel, that is, mouthings (Experiment 1) or gesture (Experiment 3), disrupts the ability to match the primary channel, that is, manual signs or speech, respectively, to the picture. We further investigate the mutual interaction of the two channels in each language: does incongruent mouthing disrupt sign comprehension to the same extent as an incongruent manual form disrupts mouthing comprehension (Experiment 2); does incongruent gesture disrupt speech comprehension to the same extent as incongruent speech disrupts gesture comprehension (Experiment 4)? Finally, we test whether the effects are modulated by language proficiency (Experiments 2 and 4) and, for the BSL experiments, hearing status (Experiments 1 and 2).

2. Making sense of the hands and mouth: British Sign Language

2.1. Experiment 1: Integration of mouthing with manual signs

2.1.1. Participants

In all, 26 native BSL signers (17 females) participated in the study in exchange for payment (age range: 18–57; mean age: 28). In all, 16 were deaf and 10 were hearing. All participants had normal or corrected-to-normal vision.

2.1.2. Materials

We used color photographs of foods as picture primes. We chose foods as a concrete semantic domain with a high degree of familiarity, for which manual signs are likely to be accompanied by mouthing in naturalistic settings and for which sign variability is expected to be low. Photographs of food items were found online and drawn from Creative Commons sources. Items were pictured against a white background. Materials consisted of 36 pictures of foods and 72 videos of a BSL signer producing signs (manual plus mouthing components) for the food items.

We recorded a native deaf BSL signer producing signs denoting foods. The manual component of each sign was accompanied by the mouthing pattern typical of standard production of the signs. The signer's hands were in his lap at video onset and offset. We constructed stimulus materials consisting of congruent and incongruent hand–mouthing

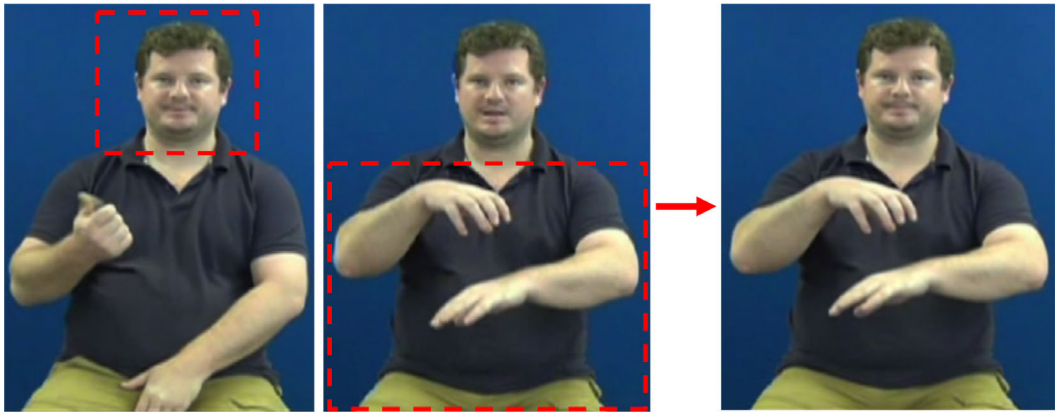

Fig. 1. Schematic representation of how the BSL video stimulus materials were created. The still frames to the left of the arrow are from the two congruent input videos (PEAS<peas>; CAKE<peas>). As represented by the dotted red lines, we take the head/face portion of one input video (showing the mouthing) and overlay it onto the body (showing the manual sign) of the other video. The still frame to the right of the arrow is from the incongruent stimulus video created through the overlay process (CAKE<peas>, i.e., CAKE produced by the hands with the mouthing "peas").

combinations using Final Cut Pro 6.0. Signs (and their accompanying mouth movements) were produced for the 36 food items, and stimulus videos were created by overlaying the face from one video onto the body of another video (see Fig. 1). In half of the videos, the manual and mouthing components were congruent, as would be typically encountered (36 videos); in the other half of the videos, hands and mouthing were incongruent (36 videos). To facilitate seamless merging of the two videos, we chose videos with minimal movement of the head and shoulders, and with signs produced below collar level. We used the same editing procedure for both congruent and incongruent stimuli so that videos did not differ in this respect. In creating incongruent combinations, we avoided pairings that were visually similar in either channel. To do this, we created a similarity matrix based on viseme similarity and formational sign parameter similarity. We assigned the onset, nucleus, and coda phonemes of each syllable of the English words appearing as mouthings to viseme categories (i.e., categories of phonemes that look the same on the lips; Fisher, 1968). Similarly, we assigned the formational parameters of the manual components of signs to different categories, based on visual similarity. We then used Excel to create a matrix that checked each possible combination of pairs for viseme category overlap and manual category overlap. The matrix returned only usable combinations (i.e., combinations that did not overlap in mouthing and manual categories) of items in cells. We also avoided pairings that were semantically similar.

2.1.3. Procedure

The experiment was conducted in a small booth containing a single computer. A name agreement phase preceded the experiment to ensure that participants associated the intended sign with each picture (important due to regional variation in BSL lexical

forms). Pictures appeared individually on the screen. After viewing a picture, participants produced their sign (and mouth pattern), which was video recorded. The main experiment commenced immediately following the name agreement phase. Participants were told that the target stimuli consisted of videos showing a male actor producing a BSL sign, and that they should pay attention to his hands only, judging whether the hand pattern in the video matched the meaning conveyed in the picture by pressing the "j" key for a match and the "f" key for a mismatch. Participants viewed the stimuli on a computer screen with a resolution of 1,024 × 768 pixels. Pictures and videos were presented on a white background in the middle of the screen. The sequence and timing of individual trials were as follows: fixation cross (displayed for 500 ms); picture (1,000 ms); stimulus video (displayed until the "f" or "j" response key was pressed); blank screen (500 ms). In practice trials only, feedback was displayed (1,500 ms) following the response key press. In the main experiment, there was a break after every 25 trials.

Practice trials preceded the experimental trials. Practice consisted of 12 trials using items that were not included in the main experiment (six match and six mismatch trials, half congruent and half incongruent in each case). In the practice trials, participants received feedback onscreen (correct/incorrect) and the experimenter was present to ensure that participants fully understood the task. The main experiment consisted of 144 trials. All participants saw the 72 videos (half hands–mouth congruent; half hands–mouth incongruent) two times: once in a hands–picture match trial (yes response) and once in a hands–picture mismatch trial (no response). Trials were presented in four blocks; each target food item appeared on the hands once per block, in one of four conditions (see Fig. 2): hands–mouthing congruent (hands–picture match); hands–mouthing incongruent (hands–picture match); hands–mouthing incongruent (hands–picture mismatch); and hands–mouthing congruent (hands–picture mismatch). The order of trials within each block was randomized for each participant.

2.1.4. Results and discussion

We first analyzed participants' production of signs (and mouthings) in the name agreement phase, excluding items on a case-by-case basis where a participant produced a substantively different manual sign than the one depicted (including fingerspelling or compound forms) or did not have a lexical sign for that item. We also evaluated mouthings (spontaneously produced on more than 95% of occurrences) and excluded items for which participants produced different mouthings than intended. For analyses, we considered accuracy (proportion correct) and trimmed correct reaction times (only including responses between 250 and 5,000 ms). No participants or items were excluded due to low accuracy (below 75% accuracy). We tested the factorial combination of hands–mouthing congruence (congruent, incongruent) and hands–picture pairing (hands–picture match, hands–picture mismatch) per trial, using mixed-effects logistic regression with crossed random effects for participants and items. To do so, we used the package lme4 (version 1.1-21: Bates, Maechler, & Bolker, 2013) running in R version 3.6.0 (R Core Team, 2013), with p values estimated using lmerTest version 3.1 (Kuznetsova, Brockhoff, & Christensen, 2017). In addition to random intercepts for participants and items, we also

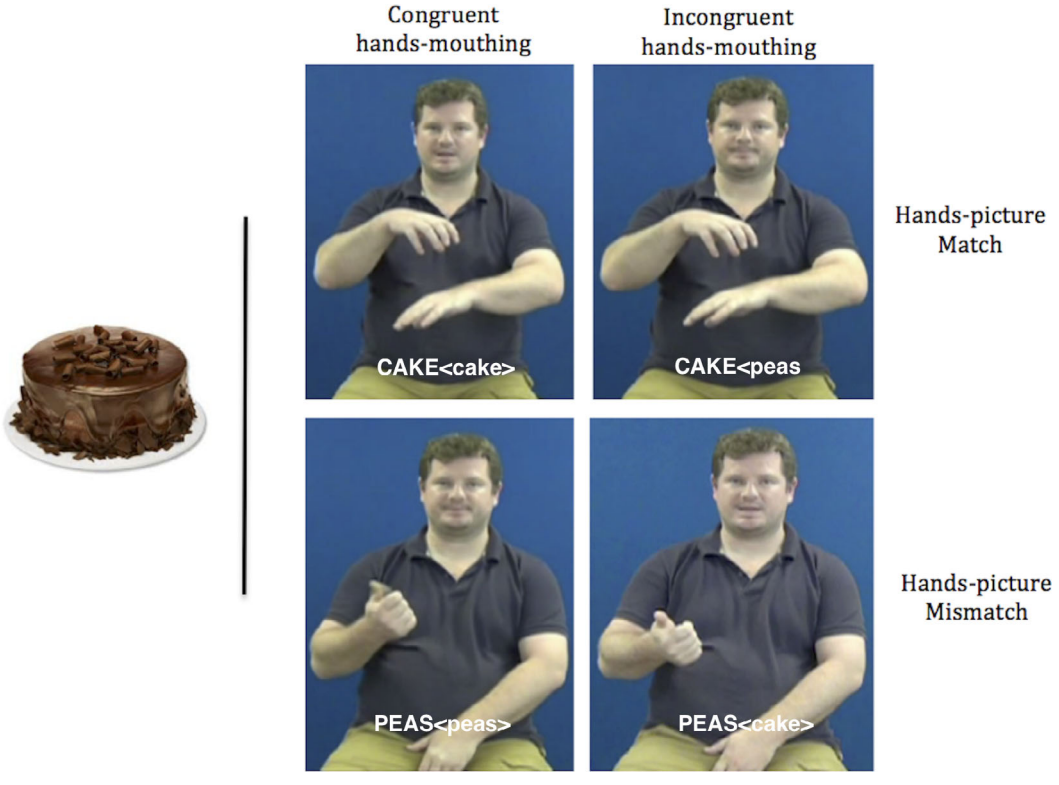

Fig. 2. Illustration of the four conditions in the picture–video matching task in Experiment 1. To the left of the vertical black bar is a picture of the food item 'cake'. To the right of the vertical black bar are still frames of videos exemplifying the four types of trials, indicating the hands–mouthing composition of each sign as "MANUAL<mouthing>." The top two still frames correspond to hands–picture match trials ("yes" responses). The bottom two still frames correspond to hands–picture mismatch trials ("no" responses).

included random by-participant slopes for hands–mouthing congruence and hands–picture match. To test whether interactions were warranted, we fit models including only the main effects and retained the more complex model only when the likelihood ratio test was significant. We also included the between-participants factor of deaf status (Deaf/Hearing) and assessed its interaction with stimulus type (Congruent/Incongruent) and hands–picture pairing (Hands–Picture Match/Mismatch).

For accuracy, in a first step, we tested whether the three-way interaction between deaf status, stimulus type, and hands–picture pairing was warranted, comparing a model with this interaction to one without it (including the two-way interactions and main effects). The three-way interaction was not warranted ($\chi^2 < 1$), but there was a significant interaction between congruence and hands–picture pairing (β [Incongruent, Hands–Picture Match–Mismatch] = −0.023, SE = 0.009, t = −2.400, p = .016). To gain a greater understanding of this interaction, we carried out separate analyses for Hands–Picture Match (yes responses) and Hands–Picture Mismatch (no responses). For yes responses, the main effect of congruence was reliable (β_{diff} = 0.042, SE = 0.019,

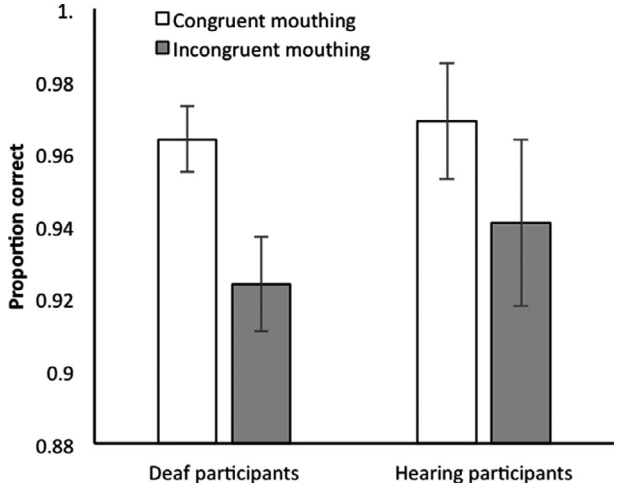

Fig. 3. Proportion correct when participants were asked to indicate whether the hands in the video matched a preceding picture, as a function of hand–mouth congruence and participant group. Error bars represent 95% confidence interval of the cell mean taking participant and item variability into account.

$t = -2.73$, $p = .009$): Participants were more accurate in judging the match between the BSL sign and the picture when the mouthing in the video was congruent compared to when it was in congruent (see Fig. 3). Neither the main effect of group nor the interaction with congruence were reliable (both $|t| < 1$). For no responses, there were no significant effects of congruence or group (all $|t| < 1.25$, $p > .22$).

For response times, we found no main effects or interactions involving any of the variables tested (main effect of Group β [Hearing—Deaf] = 115 ms, $SE = 86$, $t = 1.36$, $p = .193$; all other $|t| < 1.1$, $p > .29$), perhaps due to considerable variability in response times.

Even though participants were instructed to attend only to the hands, incongruent mouth patterns brought an accuracy cost in the task. These findings provide the first evidence that native signers habitually attend to mouth patterns accompanying signs. We found no difference between deaf and hearing native signers, though language experience may differ for mouth movements, most obviously with respect to speechreading English and having a phonological representation based on an auditory signal. Incongruent mouthing was equally disrupting to both groups.

2.2. Experiment 2: Bidirectional integration of mouthing and manual sign

We now turn to assessing the relative contribution of information from the hands and mouth when both cues are relevant: How are sign comprehenders affected by incongruence in one channel or the other? We also investigate how hearing status as well as language proficiency might modulate the use of these cues. Thus, we included five groups: (a) deaf native BSL signers; (b) hearing native BSL signers; (c) deaf BSL signers who are fluent late learners; (d) hearing BSL signers who are fluent but not native; and (e) a group

of intermediate-advanced BSL learners who are not fluent and began learning BSL only in adulthood. This last group is particularly interesting when it comes to mouth movements as it has been widely reported that late learners exhibit more effects of the surrounding spoken language (e.g., syntactic structures, under-use of bodily enactment, and classifier constructions, Ferrara & Nilsson, 2017). If this is the case, we may see a greater influence of mouthings, as supporting more "spoken-focused" processing of sign language.

2.2.1. Participants

Participants were 12 deaf native BSL signers (8 females; age range 19–46; mean age 29.2) and 8 hearing native BSL signers (3 females; age range 21–47; mean age 32.7). All deaf native participants considered BSL to be their primary language of communication. Hearing native participants used BSL on a daily basis, often professionally as interpreters. There were a total of 28 participants who were fluent, but non-native BSL signers (all self-rated as "fluent," selecting values of 6 [“fluent”] or 7 [“native/native-like”] on a 1–7 scale and had been using BSL for at least 10 years). Of these, 13 were deaf signers (10 females; age range 18–51; mean age 32.1) and 15 were hearing signers (11 females; age range 28–47; mean age 35.1). Finally, there were eight hearing signers who were at an intermediate-advanced level of BSL signing (6 females; age range 24–44; mean age 30.7). These signers had all surpassed BSL Level 2 certification (able to deal with most language tasks with a variety of BSL signers) and gave a self-rating of 4–5 on a 1–7 scale of proficiency. All participants had normal or corrected-to-normal vision.

2.2.2. Materials

In addition to the materials from Experiment 1, an additional set of object pictures and signs (36) referring to tools and other human-made artifacts was added. The tools/artifact signs were filmed and edited under the same constraints as the food signs from Experiment 1 (a total of 72 videos), and comparable photographic stimuli were prepared. Videos with incongruent combinations of hands and mouthing (36 videos) were always created within category (foods paired with foods, and artifacts with artifacts) and combined to be minimally semantically related within the category. The final set of materials for Experiment 2 comprised 36 food pictures and 72 food videos (half congruent and half incongruent) and 32 tool pictures and 64 tool videos (half congruent and half incongruent). Four tools (videos and pictures) were excluded due to problems with target picture identification or video stimulus creation.

We created a master set of 544 possible trials. Each of the incongruent videos (36 foods; 32 tools) appeared once in a Hands Match trial (the hands in the video matched the picture), once in a Mouth Match trial (the mouthing in the video matched the picture), and twice as filler trials (where neither the hands nor the mouthing matched the picture). Each of the congruent videos (36 foods; 32 tools) appeared twice in a Both Match trial (both hands and mouthing matched the picture) and twice in filler trials (neither hands nor mouthing matched the picture; see Fig. 4). The master set was divided into two lists (List A and List B). Half the participants got List A and half got List B (randomly assigned); each participant saw only half of the possible trial set (272 trials per participant).

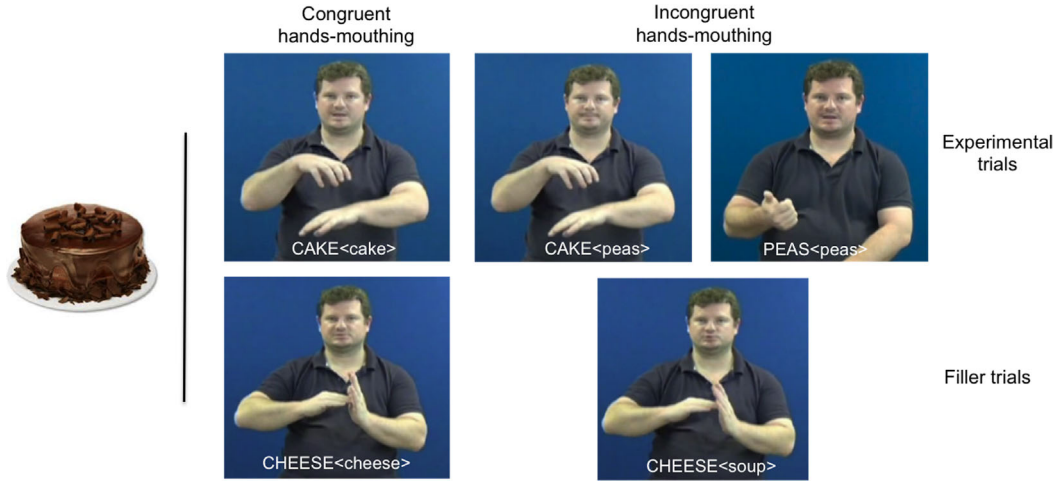

Fig. 4. Illustration of the different conditions in the picture–video matching task in Experiment 2. To the left of the vertical black bar is a photograph portraying the food item *cake*. To the right of the vertical black bar are still frames of videos exemplifying the five conditions. The top three still frames exemplify the experimental conditions (Both Match, Hands Match, Mouth Match); the bottom two still frames show filler trials.

2.2.3. Procedure

The experiment was conducted in the same manner as Experiment 1; some participants were tested off-site, in an undisturbed location and using a laptop with screen situated to provide a comparable viewing angle to the desktop PC used for other participants. As in Experiment 1, a name agreement phase preceded the experiment to ensure that participants associated the intended sign with each picture. Pictures appeared individually on the screen. After viewing a picture, participants produced their sign (and mouth pattern), which was video recorded. The experiment commenced immediately following the name agreement phase, starting with practice trials. Practice began with four demonstration examples, the first two including videos with congruent hands and mouth pattern, and the second two highlighting the incongruence and emphasizing that a "no" response should only be made when neither hands nor mouth matched the picture. There were then 20 practice trials using items that were not included in the main experiment, and participants received visual feedback about accuracy after each practice trial. The main experiment trials followed immediately thereafter but did not include feedback. Each participant saw 272 trials and had the opportunity to take a self-paced break if needed every 40 trials. The order of trials was randomized, with foods and tools mixed randomly, for each participant.

2.2.4. Results and discussion

As in Experiment 1, we analyzed the name agreement data first, excluding items on a case-by-case basis when participants produced a different sign (or no sign) or mouthing. All participants in all groups reliably produced mouthings for nearly all signs despite not being explicitly instructed to do so.

We then carried out analyses on experimental trials only, again using the same type of approach as in previous experiments. We began by testing an omnibus model incorporating match type (Both Match, Hands Match, Mouth Match) × Group, with five levels of group: deaf native; hearing native; deaf fluent non-native; hearing fluent non-native; hearing non-fluent (we were unable to test hearing status factorially because we did not have a sample of deaf non-fluent signers).[3]

For proportion correct, the model containing an interaction was a significant improvement on a comparable model with main effects only ($\chi^2(4) = 13.249$, $p = .010$). This interaction was driven by a difference between the non-fluent group and all others: There was no interaction in a model excluding the non-fluent group (comparing models with and without the interaction for the four remaining groups, $\chi^2(3) = 2.991$, $p = 0.393$). To better understand the interaction (see Fig. 5), we carried out analyses of simple main effects of match type, separately for hearing non-fluent signers, and then for all the others combined. Combining the latter groups is warranted for this analysis because the model comparison indicated no interaction between group and match type, once non-fluent participants were removed from the sample. For these simple main effects analyses, we fit the same types of models but treated Hands Match as the reference condition. Among hearing non-fluent signers, the difference between Both Match and Hands Match was not significant (β =0 .025, $SE = 0.036$, $t < 1$), but there was a significant difference between Mouth Match and Hands Match ($\beta = -0.179$, $SE = 0.071$, $t = 2.931$, $p = .008$). Instead, among fluent signers there was a difference between Both Match and Hands Match ($\beta = 0.046$, SE =0 .019, $t = 2.190$, $p = .031$), and between Mouth Match and Hands Match ($\beta = -0.075$, $SE = 0.023$, $t = 2.055$ $p = .044$). While fluent signers showed a benefit of mouth congruence (Both Match > Hands Match), non-fluent signers did not.

For response times, the model including the interaction between group and match type was not a significant improvement over the model with main effects only ($\chi^2(4) < 1$). There was a main effect of group (illustrated by a significant increase in RT for advanced non-fluent signers compared to the reference group: native deaf signers ($\beta = 251$ ms, $SE = 78$, $t = 3.18$, $p = .002$), but no significant effect of match type ($|t| < 1.1$, $p > .30$ for Both Match vs. Hands Match, and for Both Match vs. Mouth Match). Again, we removed advanced non-fluent signers from the dataset and fit one final model (Group + Match Type); all comparisons involving group and match type were non-significant ($|t| < 1.15$, $p > .25$). Overall, this reveals that non-fluent signers were slower on the task regardless of stimulus type; among fluent signers we found no response time differences based on age of learning BSL or deaf status.

For fluent signers, response accuracy was affected significantly when only the hands matched, showing again that the mouth matters in sign comprehension; accuracy was significantly worse again when only the mouth matched. Thus, for fluent signers, regardless of hearing status or age of learning BSL, the findings reveal a strong imbalance between hands and mouth, with the former—the primary channel—more relevant. The advanced but non-fluent group showed a very different pattern: When only the hands matched, non-fluent signers were no worse in their response than when both cues matched. When only the mouth matched, non-fluent signers fared particularly poorly in terms of accuracy.

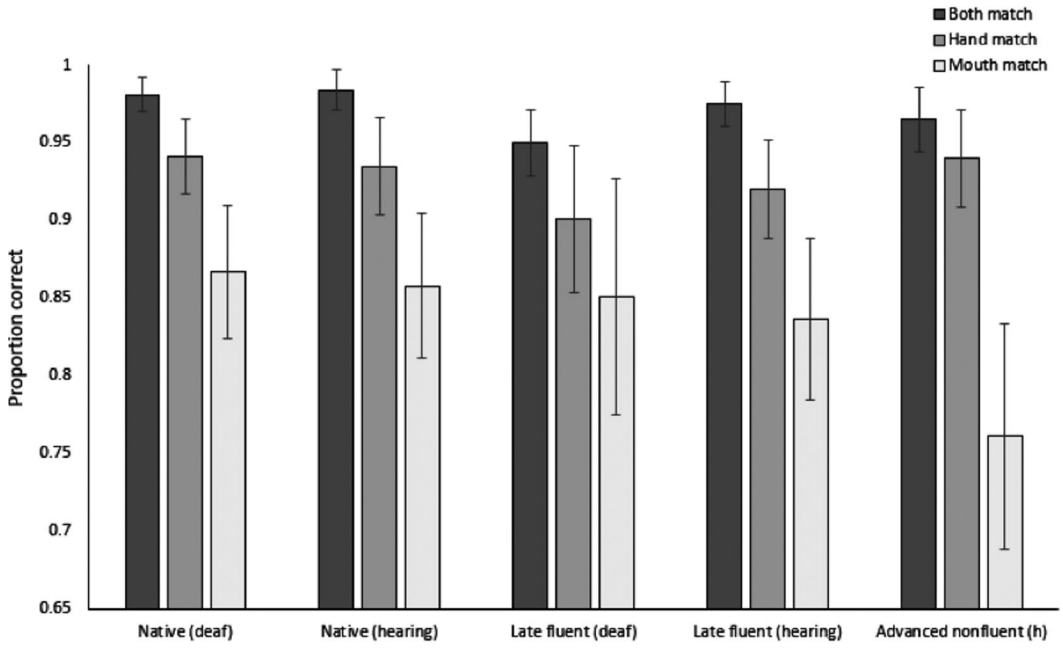

Fig. 5. Proportion correct as a function of group and BSL hand–mouth match condition. Error bars represent 95% confidence interval of the cell mean taking participant and item variability into account.

Although signers in this group do produce mouth patterns themselves, the results suggest that they rely overwhelmingly on manual configuration during comprehension.

3. Making sense of the hands and mouth: English

3.1. Experiment 3: Integration of gesture with audio-visual speech

3.1.1. Participants

In all, 71 first-year psychology students (59 females; 12 males) at University College London participated in the study for course credit as part of a laboratory session (age range 17–21; mean age 18.61). All participants were native speakers of English (22 were fluent in another language as well). All participants had normal or corrected-to-normal vision and normal hearing.

3.1.2. Materials

The stimulus materials for experiment 3 consisted of 30 black-and-white line drawings of objects (e.g., *ball*) and actions (e.g., *tearing* [*paper*]) and 60 video clips of a male actor producing object or action words accompanied by an iconic gesture. Objects and actions were chosen as semantic categories because they are both easily pictureable and

gestureable. The line drawings were comprised from Druks and Masterson (2000) and images from Creative Commons sources. All pictures had a resolution of 350 × 350 pixels. We used Adobe Photoshop to modify individual images according to the needs of the study and in order to make the set stylistically uniform. The picture set comprised 18 objects and 12 actions. The videos included 36 object (18 congruent; 18 incongruent) and 24 action (12 congruent; 12 incongruent) videos, comprised of 18 different object and 12 different action target items (30 target items total), and corresponding to the stimulus pictures of objects and actions. All video clips were 720 × 576 pixels in AVI format.

To produce the videos, we recorded an actor producing words denoting objects and actions accompanied by an iconic co-speech gesture representing features of the object or action. For objects, the gestures either depicted a movement associated with the object (e.g., a loosely closed hand twisting back and forth to represent *screwdriver*) or outlined the object's shape (e.g., the hands tracing a circle to represent *ball*). For actions, gestures depicted the manual manipulation of the object involved (e.g., holding an iron and moving it back and forth to represent *ironing*) or represented the bodily movement involved in the action (e.g., moving open hands away from the body to represent *pushing*). In all videos, the actor's hands were in his lap at video onset and returned to his lap after production of each item. Eight native speakers of English viewed the videos of speech–gesture combinations, rating each one on a 0–5 scale (0 ="gesture does not reflect the speech at all" and 5 = "gesture reflects the speech very highly"). Only items that received a rating of 4–5 were included in the experiment.

We constructed stimulus materials consisting of congruent and incongruent speech–gesture combinations using the video editing software Final Cut Pro 6.0. In half of the videos, speech and gesture were congruent and expressed the same meaning (e.g., *pushing* in speech accompanied by a pushing gesture); in the other half of the videos, speech and gesture were incongruent and did not express the same or a similar meaning (e.g., *pushing* in speech accompanied by a tearing gesture). Creation of the incongruent speech–gesture pairings was subject to the following additional constraints: We chose words of the same syllable length, but whose onset phonemes differed (e.g., combining *pushing*/*tearing* and *ball*/*can*, but not *pushing*/*poking* or *car*/*can*). In addition, we avoided combining words whose accompanying gestures exhibited form similarity in movement and/or handshape (e.g., *poking*/*punching*). Finally, as with the BSL materials, we chose videos with minimal movement of the head and shoulders, and with gestures produced below collar level to facilitate seamless merging of the two videos.

We created stimulus videos in the same way as for the BSL hands–mouthing materials by overlaying the face from one video onto the body of another video (see Fig. 6). We retained only the audio from the face video (top), deleting the audio track from the body video (bottom). In this way, we could mismatch speech and gesture while maintaining congruence between the heard word and the visible movements of the face/mouth. In overlaying the two videos, we took care that the timing of speech and gesture onset looked natural, by aligning speech onset for both clips. As a result, gesture onsets slightly preceded speech onset as occurs in natural communication (Morrel-Samuels & Krauss, 1992; Schegloff, 1984). Again, both congruent and incongruent stimuli were edited in this

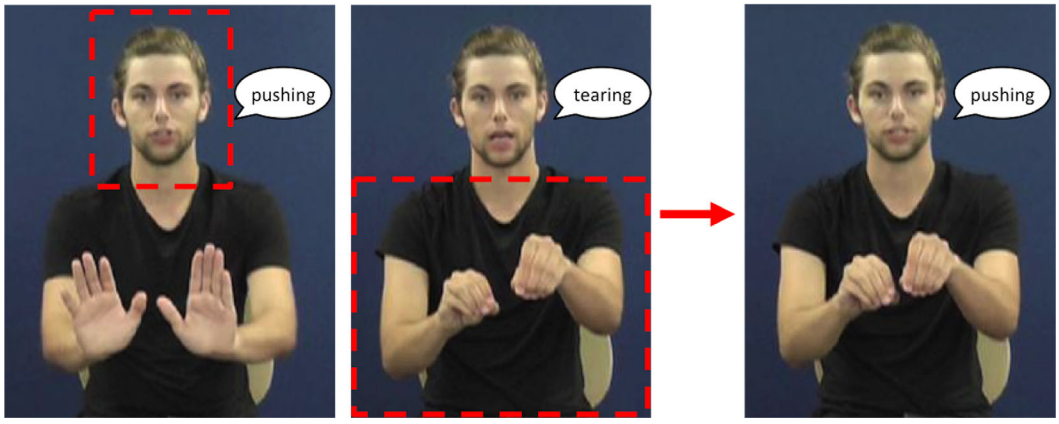

Fig. 6. Schematic representation of how the speech–gesture video stimulus materials were created, with a final incongruent video of "pushing" in speech combined with a co-speech gesture depicting "tearing." The head/face portion of one input video together with the audio of the spoken word is overlaid onto the body, depicting the gesture, of the other video.

way. The edited videos were shown to eight native speakers of English naïve to the manipulation. Beyond the mismatch between speech and gesture, none of these volunteers reported noticing anything otherwise remarkable about the videos and were unaware that the videos had been edited.

3.1.3. Procedure

For the experiments in English, a picture familiarization phase preceded the experiment to ensure that participants associated the intended word with each picture. Pictures appeared individually on the screen. After viewing a picture, participants pressed a key to display the intended word. Object and action pictures were shown in blocks, in the same order as in the main experiment (see below). Task instructions were presented on the screen after the familiarization phase was completed, followed by practice trials to ensure that participants understood the task.

The experiment was conducted in a large computer lab. Each participant was seated in front of a computer screen and used a keyboard to log responses. The participants' task was to decide whether the object or action mentioned in an audio-visual stimulus matched the object or action in a previously presented picture. Participants were told that the target stimuli consisted of videos showing a male actor saying a word aloud and that the speech in the video would either match or mismatch the picture seen; the presence of gesture in the videos was not explicitly mentioned. Participants judged whether the speech in the video matched the meaning conveyed in the picture by pressing the "j" key for a speech–picture match and the "f" key for a speech–picture mismatch.

Practice consisted of 12 trials (speech matched the picture in six trials; speech mismatched the picture in six trials), and participants received feedback onscreen (correct/

incorrect) following their response. The main experiment was the same as in practice, but with no feedback. There were four experimental conditions representing the factorial combination of speech–gesture congruence (Congruent, Incongruent) and speech–picture pairing (Speech–Picture Match, "yes" response; Speech–Picture Mismatch, "no" response). The main experiment consisted of a total of 120 trials (72 object trials; 48 action trials). Each video and each picture appeared twice, once in a speech-picture match trial and once in a speech-picture mismatch trial (as in Experiment 1, see Fig. 2). Object and action trials were presented in blocks with the order of blocks (action–object; object–action) assigned by odd/even participant number. When one block was completed, task instructions and practice trials were repeated for the second block. The order of trials within object/action blocks was randomized.

Participants viewed the stimuli on a computer screen with a resolution of 1,024 × 768 pixels. Speech was presented through headphones. Pictures and videos were presented on a white background in the middle of the screen. The sequence and timing of individual trials were as follows: fixation cross (displayed for 500 ms); picture (1,000 ms); speech–gesture stimulus video (displayed until the "f" or "j" response key was pressed); blank screen (500 ms). In practice trials only, feedback was displayed (1,500 ms) following the response key press. In the main experiment, there was a break after every 25 trials.

3.1.4. Results and discussion

For analyses, we considered accuracy (proportion correct) and trimmed correct reaction times (only including responses between 250 and 5,000 ms). No participants or items were excluded due to low accuracy (below 75% accuracy). One participant was excluded because the audio was not working during the task. We tested the factorial combination of speech–gesture congruence (Congruent, Incongruent) and speech–picture pairing (Speech–Picture Match, Speech–Picture Mismatch)[4] per trial, using mixed-effects logistic regression with crossed random effects for participants and items. Subsequent analyses were conducted using the same types of models as in Experiment 1.

For proportion correct (see Fig. 7a), the model containing the interaction was a significant improvement over the model without it ($\chi^2(1) = 17.1$, $p < .001$) so we retained the interaction in the final model. The main effect of speech–picture match was not significant ($\beta_{diff} = 0.003$, $SE = 0.007$, $t = 0.476$), while the main effect of congruence was reliable ($\beta_{diff} = 0.050$, standard error of the estimate $= 0.010$, $t = 4.98$, $p < .001$). They were modulated by a significant interaction ($\beta_{diff} = 0.040$, $SE = 0.010$, $t = 4.138$, $p < .001$). To understand the interaction, we tested simple main effects of picture–gesture congruence using separate models. When the speech matched the picture ("yes" trials), congruent stimuli elicited significantly more accurate responses than incongruent ones ($\beta_{diff} = 0.050$, $SE = 0.013$, $t = -3.983$, $p < .001$). However, when speech mismatched the picture ("no" trials), there was no difference ($\beta_{diff} = 0.010$, $SE = 0.009$, $t = -1.101$, $p = .275$).

For reaction time (see Fig. 7b), the model containing the interaction was not significantly better ($\chi^2(1) = 0.38$, $p = .5375$), so the final model contained only main effects. The main effect of speech–picture match was significant ($\beta_{diff} = 135$ ms, $SE = 13$, $t = 10.722$, $p < .001$): Mismatch trials were slower than match trials. The main effect of

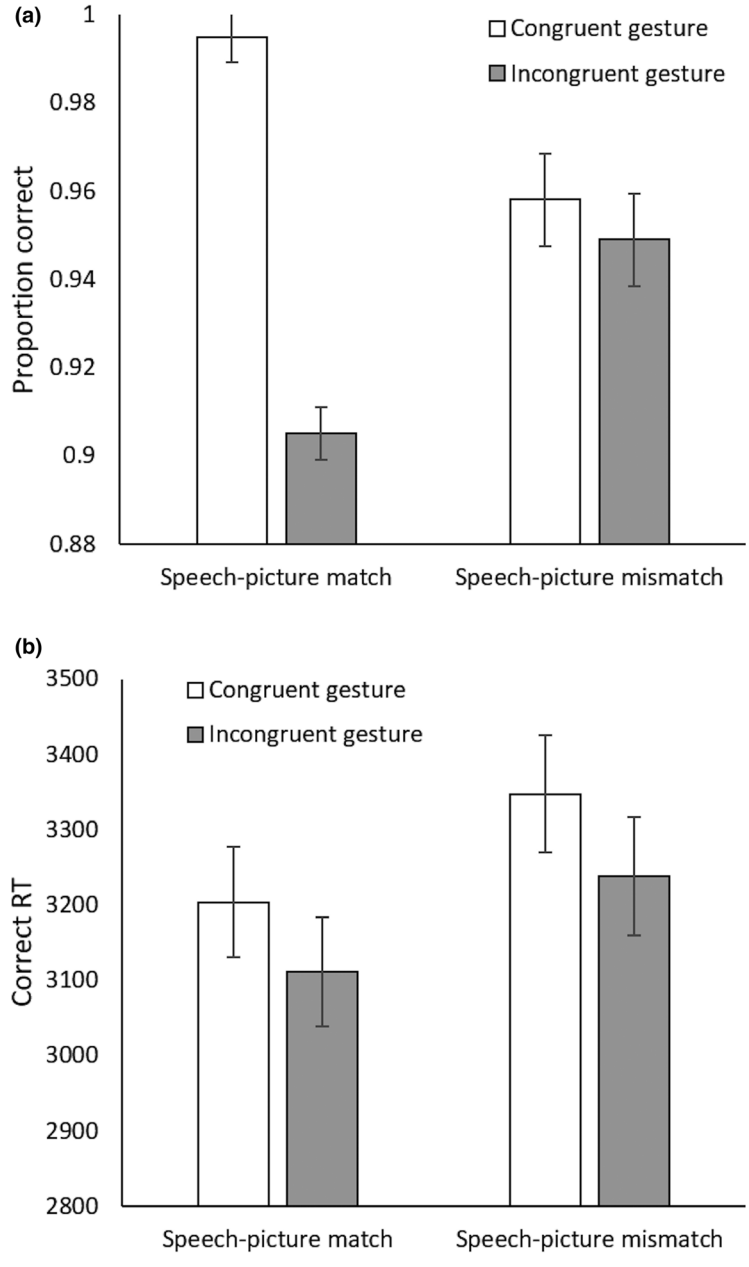

Fig. 7. Proportion correct (a) and trimmed correct reaction time (b) when participants were asked to indicate whether speech in the video matched a preceding picture, as a function of speech–gesture congruence and speech–picture pairing. Error bars represent one standard error of the estimated cell mean taking participant and item variability into account.

speech–gesture congruence was also significant (β_{diff} = 102 ms, *SE* = 17, $t = -5.818$, $p < .001$): Congruent speech–gesture pairings elicited slower responses regardless of whether the speech matched the picture or not.

Replicating findings by Kelly et al. (2010), we found an effect of speech–gesture congruence on task performance: Participants were less accurate for incongruent speech–gesture pairings. Participants were unable to ignore an incongruent gesture even though it was irrelevant to the task. Our results are also in line with behavioural findings for native speakers of Dutch reported in Drijvers and Özyürek (2018). Interestingly, for RTs, we found that overall congruent speech–gesture pairings were answered slower. The reason for this is unclear.

3.2. Experiment 4: Mutual interaction between audio-visual speech and gesture

The previous experiment established that language comprehenders integrate meaningful cues from the hands with speech. This happens even when gesture is not explicitly relevant to the task at hand. We now turn to assessing the relative contribution of speech and gesture when both cues are relevant. We further ask whether the weight of these cues in processing differs between native and non-native speakers of English.

3.2.1. Participants

Participants were visitors to the London Science Museum's Live Science area. Data were collected from a total of 188 participants, of whom 119 were native speakers of English (78 females, 41 males; age range 16–75; mean age 32.1) and 69 were non-native speakers of English (40 females, 28 males, one declined to answer; age range 16–57; mean age 29.2). Non-native speakers were asked to rate their proficiency in English on a scale of 1–5 (with 1 = "not very good" and 5 = "near native"; mean rating = 4.03). (Only visitors who rated their English proficiency as 3–5 were assigned to the present task; participants who rated their proficiency as 1 or 2 were assigned to a different experiment running concurrently in the Live Science area.) Non-native speakers reported the age of first exposure to English as 9.6 on average (range 1–23). All participants had normal or corrected to normal vision and normal hearing.

3.2.2. Materials

Stimulus materials for Experiment 4 consisted of 48 color photographs of people performing actions (e.g., *tearing*, *driving*, *pouring*) and 96 video clips of an actor producing speech–gesture combinations denoting these target actions. The set of color photographs was comprised in part from Speechmark color cards of actions (from the *Early actions* and *Familiar verbs* sets) and in part from photos available online for which reuse for noncommercial purposes was permitted. Six native speakers were asked to verify whether each picture matched the target speech, and items that did not meet this criterion were not used in the study. Larger pictures were reduced to fit into a 600 × 600 pixel square but retained their original height:width ratio; smaller pictures were kept in their original size (all greater than 340 pixels in the smallest dimension).

As for Experiment 3, we recorded a male actor producing action words accompanied by an iconic gesture representing the action, and then edited the video clips in the same way as for Experiment 3 such that target videos (96) consisted of congruent (48) and incongruent (48) speech–gesture combinations. The final set of videos contains 48 different actions that correspond to the actions represented in the picture stimuli. All video clips were 720 × 576 pixels in AVI format.

We created a master set of 384 possible trials. Each of the 96 videos appeared four times in the complete design across participants. For the 48 incongruent videos, each one appeared in one Speech Match trial (the spoken word matched the picture while the gesture mismatched), one Gesture Match trial (the gesture produced matched the picture while the word did not), and two filler trials (neither the word nor the gesture matched the picture), thus balancing the number of "yes" and "no" responses in the set as a whole. The 48 congruent videos also appeared four times: twice matching the picture (Both Match) and twice as fillers. Pictures for filler trials were assigned pseudo-randomly, using a reordering of the set for experimental trials with the constraint that neither speech nor gesture should match the picture. In all, 12 lists were created from this master set, using pseudorandom selection of trials from the master list under the following constraints. Each list contained 32 trials: 16 experimental and 16 fillers. Experimental trials in each list comprised eight congruent speech–gesture trials (Both Match condition) and eight incongruent speech–gesture trials (four in Speech Match and four in Gesture Match condition) (as in Experiment 2, see Fig. 4). Filler trials in each list comprised an equal number of congruent (8) and incongruent (8) video clips.

3.2.3. Procedure

Data were collected over the course of 18 days at the London Science Museum as part of the museum's Live Science scheme. Visitors to the Live Science area were told that researchers were conducting experiments about how we understand language and communicate with each other. After expressing interest in participating, visitors signed a consent form and were led to one of three computers in the Live Science space. Participants first filled in an online questionnaire (indicating their age, gender, native language, level of proficiency in English if English was not their native language, and other languages spoken/signed). As in the previous experiments, participants viewed the stimuli on a computer screen with a resolution of 1,024 × 768 pixels, with pictures and videos presented on a black background in the middle of the screen. Instructions appeared on screen: Participants were told they would be shown a picture of an action followed by a video; their task was to decide whether any part of the video matched the picture. Participants were told that the target videos showed an actor producing a word together with a gesture. As in Experiment 3, speech was presented through headphones. On a keyboard placed in front of them, participants pressed the "j" key if the speech and/or gesture in the video matched the pictured action, and pressed "f" if no part of the video (neither speech nor gesture) matched the action in the picture.

Practice trials preceded the main experiment. An experimenter sat next to the participant during practice to ensure that participants fully understood the task. Practice

consisted of 12 trials using items that were not included in the main experiment and participants received feedback onscreen (correct/incorrect) and from the experimenter following each response. The main experiment was the same as in practice, but with no feedback. Each participant completed 32 trials from one of the 12 lists described above. The sequence and timing of individual trials in practice and in the main experiment were the same as in Experiment 3. The order of the trials was randomized for each participant.

3.2.4. Results and discussion

We first calculated accuracy for each participant on the task as a whole. Two participants were excluded on the basis of accuracy (performing with accuracy less than 75% correct). Both these participants were non-native speakers who rated their English proficiency as 3 on the 1–5 scale. We then assessed accuracy for each item; any speech or gesture with accuracy of less than 60% in any experimental condition was removed across all conditions. Six items were excluded on this basis (playing volleyball, polishing, rubbing, stacking, screwing, tiptoeing).

As in Experiment 2, we analyzed the results using mixed-effects logistic regression with crossed random effects for participants and items, using lme4 (version 1.0-4: Bates et al., 2013) running in R version 3.0.1 (R Core Team, 2013). We tested the 3 × 2 factorial combination of picture match condition (Both Match, Speech Match, Gesture Match, with Both Match treated as the reference condition) and group (native speaker, non-native speaker) on accuracy and trimmed correct reaction time (between 250 and 5,000 ms) per trial. In addition to random intercepts for participants and items, we also included random by-participant slopes for picture match condition, and random by-item slopes for group.

For proportion correct (see Fig. 8a), the model containing the interaction was a significant improvement over the model without it ($\chi^2(2) = 18.5$, $p < .001$), so we retained the interaction in the final model. The main effect of group was not significant ($\beta = 0.004$, $SE = 0$ .015, $t = 0.331$, $p = .741$). There was a reliable main effect of picture match type (β_{diff} (Both Match—Gesture Match) = −0.120, $SE = 0.022$, $t = -5.360$, $p < .001$; β_{diff} (Both Match—Speech Match) = −0.085, $SE = 0.024$, $t = -3.555$, $p < .001$): Accuracy was lower for incongruent than for congruent videos. There was a significant interaction (β [Non-native, Gesture Match vs. Both match] = −0.169, $SE = 0.040$, $t = -4.220$, $p < .001$). To better understand this interaction, we fit additional models, comparing native to non-native speakers in each of the three conditions separately. There was no difference in accuracy between groups for either Both Match ($\beta = 0.006$, $SE = 0.009$, $t = 0.675$, $p = .501$) or Gesture Match ($\beta = 0.024$, $SE = 0.037$, $t = 0.667$, $p = .508$). However, for Speech Match, non-native speakers were significantly less accurate than native speakers ($\beta = -0.159$, $SE = 0.038$, $t = -4.173$, $p < .001$).

For reaction times (see Fig. 8b), the model containing the interaction was not a significant improvement over the model with only main effects ($\chi^2 < 1$), so we included only main effects in the final model. There was a significant effect of group ($\beta = 119$, $SE = 58$, $t = 2.052$, $p = .0417$): Native speakers responded faster overall than non-natives. There was also an effect of picture match type with both incongruent pairings slower than the congruent (Both Match) condition (β_{diff} [Gesture Match] = 288, $SE = 28$,

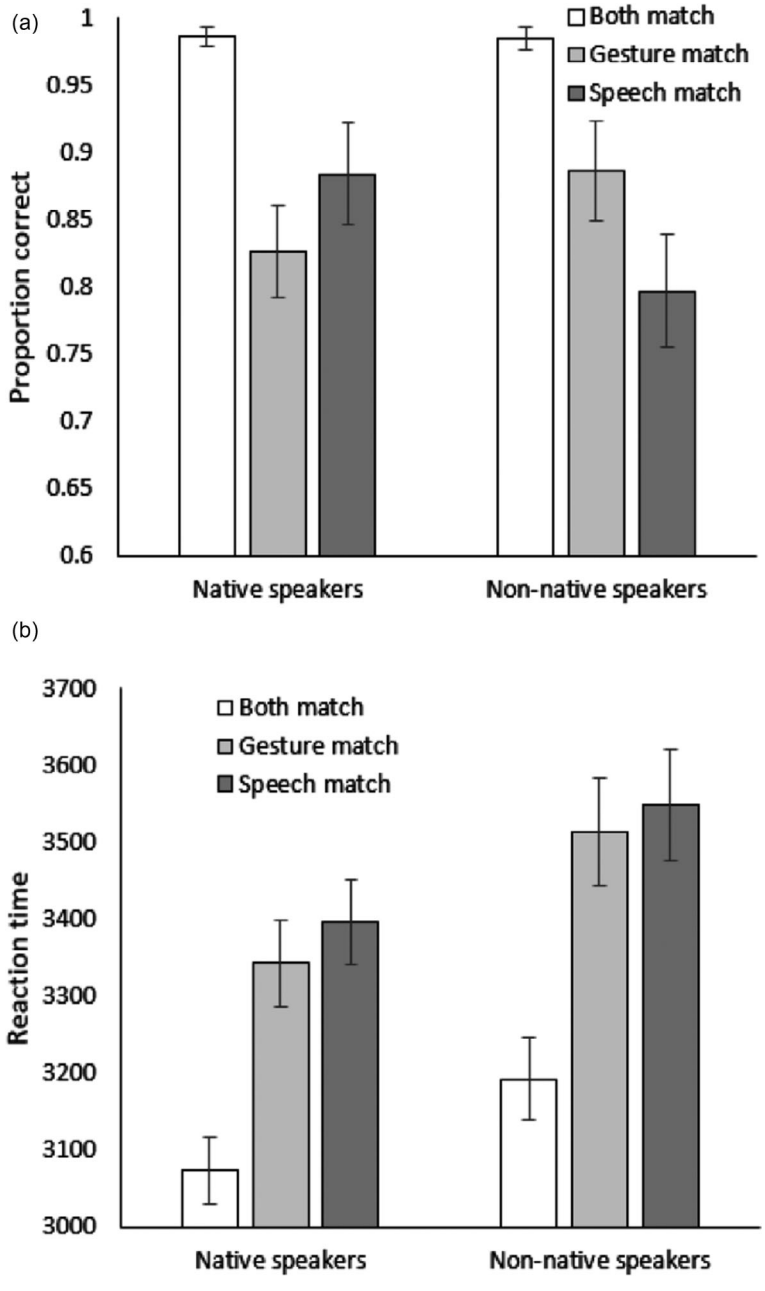

Fig. 8. Proportion correct (a) and trimmed correct reaction time (b) as a function of group and speech–gesture match condition. Error bars represent one standard error of the estimated cell mean taking participant and item variability into account.

$t = 10.4$, $p < .001$; β_{diff} [Speech Match] = 338, $SE = 26$, $t = 13.1$, $p < .001$). To test whether there were differences between the two incongruent conditions, we fit one last model, excluding the Both Match condition. The difference between these was not significant ($\beta_{diff} = 36$, $SE = 33$, $t = 1.112$, $p = .267$).

The results of Experiment 4 show that both incongruent speech and incongruent gesture have strong interfering effects on comprehension in both native and non-native speakers. For accuracy, the interaction between language background and congruence suggests that speech and gesture are used differently by native speakers and non-native speakers of English. The results for native speakers again replicate and extend findings by Kelly et al. (2010), who found no interaction between congruence and cue. For non-native speakers, in contrast, incongruent gestures created a greater cost to processing than incongruent speech. These results confirm our prediction that non-native speakers rely more on gestures than do native speakers, perhaps due to increased reliance on additional semantic information from the visual modality. For both native and non-native speakers, having both cues available and relevant afforded an advantage in response time, with significantly faster response times when speech and gesture were congruent (and both matched the picture) compared to the other conditions. In contrast to Experiment 3, we did not find longer RTs for congruent than incongruent, a finding that can be accounted for in terms of differences in task requirements: While in Experiment 3 the task focused exclusively on the speech, here participants were simultaneously monitoring both the speech and the gesture.

4. General discussion

Together, the experiments provide new insight into how different cues are used in language comprehension in both modalities, and how these effects are modulated by language proficiency. Our results provide first behavioral evidence for the integration of information from manual and non-manual articulators in a signed language. Specifically, our findings for BSL show that fluent signers exhibit an incongruence cost for both cues, but in a highly asymmetric way, making most errors when the hands mismatched the pictures, and significantly fewer errors when the mouth mismatched the picture.

The results for BSL contribute to the debate regarding the status of mouthing within the sign lexicon, and are in line with previous results, from both corpus and experimental studies, arguing that mouthings represent code-blending with the surrounding spoken language (Capek et al., 2009; Ebbinghaus & Heßmann, 2001; Giustolisi et al., 2017; Johnston et al., 2016; Vinson et al., 2010). Using a sign production task, Vinson et al. (2010) found that semantic errors did not always co-occur in the two channels in picture naming. This dissociation would not be expected if the manual and mouthing components of a sign constituted a single lexico-semantic representation. Using fMRI, Capek et al. (2009) found that brain activity for signs in which mouthing functions to disambiguate a manual minimal pair resembled activity for English speech-reading, compared to signs without mouthing. As Woll and MacSweeney (2016, p. 254) state, these findings suggest that

"the brain does appear to care about the status of mouthings used in signed languages." In a comprehension task, Giustolisi et al. (2017) asked participants to decide whether a LIS (Italian Sign Language) sign matched a following printed Italian word in meaning. Incongruent conditions used the Italian translation equivalent of a LIS minimal pair, with minimal pairs distinguished on five dimensions: the four manual formational parameters (handshape, location, movement, and orientation) and mouthing. Task performance was worst for the minimal pairs distinguished by mouthing (i.e., where the manual form is the same, and only mouthing differs, as in AUNT/BATTERY in BSL or ROME/IRON in LIS), suggesting that signers pay closer attention to the more important manual components of signs. This is consistent with findings from the present study showing an asymmetry in the degree to which information from the hands and mouth affects comprehension.

Our findings additionally suggest that attention to mouthing in comprehension is not modulated by whether the English phonological representation is based on acoustic input (as it is for hearing signers). We found no difference between native/fluent deaf and hearing signers in the nature of integration of information from the code-blended mouthing element. Mouthing is thus not used differently depending on having a phonological representation based on an auditory signal or experience with speechreading English. A very different pattern of results was found for the non-fluent signers. This group showed no significant difference between the condition where both hands and mouth matched the picture and the condition where only the hands matched. It seems that non-fluent signers are insensitive to mouth mismatches as long as the hands are semantically consistent with the picture. When only the mouth matched, non-fluent signers were much worse than the other groups. Thus, rather than non-fluent signers relying more on information from the mouth in comprehension, as information from a second language—their "source" language English—it seems that non-fluent signers use the mouth less, at least under these task constraints. The different pattern suggests that non-fluent signers do not yet sufficiently divide their attention to use these cues successfully in comprehension. What seems crucial is extensive experience with mouth movements as they occur in BSL.

Given their widespread presence (for BSL, Sutton-Spence, 2007; see also Crasborn et al., 2008; Johnston et al., 2016), mouthings are an additional reliable cue to meaning. As such, mouthings have been argued to play an important role in facilitating understanding (Sáfár et al., 2015; Stamp, 2016; Stamp et al., 2016). Stamp (2016) identifies mouthing as an accommodation strategy and highlights signers' frequent exposure and adaptability to variation, including regional lexical variation, age-based differences between older and younger signers, and differences due to large variability in acquisition patterns (as only a small percentage of deaf individuals acquire sign language from birth from parental input). Highly proficient signers may have more experience communicating across a range of different contexts and environments, and thus have more experience integrating mouthing as a cue to meaning in comprehension. In addition, recent studies on mouthing highlight the extent to which these mouth movements are an integral part of signed language (Bank et al., 2016; van de Sande & Crasborn, 2009). Mouthings reflect the constant language contact situation with the surrounding spoken language, and rather

than disappearing from signing (Hohenberger & Happ, 2001), there is evidence that younger signers use more mouthings than older signers and for functionally more sophisticated purposes (prosodic, grammatical, and stylistic; Boyes-Braem, 2001; Mohr, 2012). While mouthings clearly represent phonological information from a different language (Vinson et al., 2010), they have become intimately integrated into language use and linguistic structure. Mohr (2012) suggests that mouthings may be less like code-blends (reflecting signers' constant co-activation of two languages, Emmorey, Borinstein, & Thompson, 2005) and more like borrowed elements. Thinking of mouthings as elements more akin to loanwords that have become integrated into the linguistic system may also allow for a more straightforward account of different patterns of mouthing activity across sign languages. In support of this, Adam (2013) found different patterns of mouthing use in Australian Sign Language (Auslan) and Australian Irish Sign Language (AISL), which share mouthings based on English, in individual bilingual signers.

Nevertheless, as a cue to meaning in comprehension, the findings from BSL show that information from the mouth is clearly secondary to the information conveyed on the hands. This is in contrast to our findings from spoken language. For native speakers of English, we found that the influence of gesture on speech comprehension and of speech on gesture comprehension was relatively comparable; that is, incongruence between speech and gesture was as disruptive when speech matched the picture as when gesture matched the picture. We thus replicate findings by Kelly et al. (2010) in a different variety and community of users of English and extend the findings in an important way. The fact that we obtained these results with materials using audio-visual speech—that is, mouth movements were visible and congruent with heard speech—offers compelling support for the automatic uptake of information from gesture when gesture accompanies speech. In particular, our results rule out that previous findings by Kelly et al. (2010) reflected greater attention to co-speech gesture due to (a) an inaccessibility of visible mouth movements, and (b) a forced visual focus on the gesture as a result of seeing only the speaker's torso (supported also by findings by Drijvers & Özyürek, 2018 for speakers of Dutch).

Drijvers and Özyürek (2018) used a cued-recall task in which participants identified the spoken word from a set of four alternatives, including a phonological and a semantic competitor word. Consistent with the authors' interest in speech perception, and the potential support lent by concomitant visual cues, the cued-recall task focuses primarily on recognition of a spoken word, hence likely boosting phonological awareness. The picture–video matching task used in the present study requires a concept to be activated and recognized in speech, investigating the integration of information from speech and gesture on a semantic level more directly. Our use of a picture instead of a video prime in this paradigm (to minimize the visual and temporal overlap between the prime and the target) thus further buttresses the claim that we cannot help but pay attention to gesture in language comprehension. If a mismatching concept is encountered in gesture, the effect is detrimental to task performance: The mismatching gesture cannot be ignored.

This effect was particularly pronounced for non-native speakers of English. Non-native speakers exhibited an even greater cost when gestures were incongruent with speech,

suggesting that they rely on gesture to support and confirm the information received from speech. This is in line with previous studies suggesting that gesture processing may be especially useful when speech processing is difficult or problematic (Drijvers & Özyürek, 2017, 2018; Obermeier et al., 2012). Obermeier et al. (2012) looked at situations where hearing is difficult because of a noisy environment (as with babble noise in the background) or a hearing impairment, and Drijvers and Özyürek (2017, 2018, 2019) used a vocoder to create different levels of degraded speech, showing increasing usefulness of gesture cues as clarity of speech decreases. Thus, gesture can support comprehension by filling in information when speech is less intelligible for some reason. In our study with non-native speakers of English, speech processing is problematic as a result of reduced language proficiency. As the findings by Drijvers and Özyürek (2018, 2019) show, however, information from gesture—and indeed information from visible speech (Drijvers & Özyürek, 2019)—is beneficial only when sufficient auditory cues are available for processing speech. Drijvers and Özyürek (2018) observed an increased N400 amplitude in clear speech for incongruent compared to congruent gestures for both native and non-native speakers, but found the same difference in degraded speech only for native speakers. This suggests that non-native speakers cannot use the semantic cues from gesture when auditory cues are too difficult to resolve.

Recently, Özer and Göksun (2019) addressed the possibility that speech–gesture integration may be modulated by cross-cultural differences in gesture use. Using the same paradigm and materials (including an action video prime and speech–gesture videos showing the torso only) as Kelly et al. (2010) to investigate bidirectional integration of speech and gesture (as in our Experiment 4), Özer and Göksun (2019) likewise found that incongruent speech–gesture led to significantly higher error rates in Turkish participants compared to congruent speech–gesture pairings. However, in contrast to Kelly et al. (2010) and to the present findings, they found that participants made significantly more errors when only speech matched compared to when only gesture matched the action video prime. The authors consider whether the similarity of the action prime and target gesture may have led to the asymmetric relationship between speech and gesture, that is, that mismatching gesture would hinder comprehension more than mismatching speech. In the present study, we used a picture-matching task precisely to avoid the high degree of isomorphism between the prime and target videos and to ensure that the task would require conceptual access and semantic processing rather than simply a visual mapping between the actions presented in the videos. The results of the present study suggest that the isomorphism between actions and gestures does not explain Özer and Göksun (2019)'s finding of higher error rates when gestures mismatched.

What is interesting to consider is Özer and Göksun (2019)'s suggestion that Turkish speakers may rely more on gesture due to language-specific differences in the use of gestures. Azar, Backus, and Özyürek (2020) found a higher rate of iconic gesture use in speakers of Turkish (a pro-drop, verb-framed language) compared to speakers of Dutch (a non-pro-drop, satellite-framed language like English). These typological characteristics mean that many utterances in Turkish may consist of just a verb (with omitted arguments) and an iconic gesture, leading to a greater focus on gestures by Turkish speakers.

Support for this argument comes from language development: Furman, Küntay, and Özyürek (2014) found that iconic gestures appear earlier in production in Turkish-speaking children (from 19 months) compared to English-speaking children (about 24 months).

The findings by Özer and Göksun (2019) for Turkish speakers mirror our findings for non-native speakers of English. A greater reliance on gesture as a cue to meaning may be due to a lack of language proficiency, but it may also come about due to language-specific differences in information packaging. Social, situational, and contextual factors have also been shown to modulate the nature of processing different cues (e.g., Holler et al., 2014, 2015), and brain network involvement and strength of integration has been shown to vary according to task demands (e.g., Obermeier et al., 2015; Yang, Andric, & Mathew, 2015; see Cocks, Byrne, Pritchard, Morgan, & Dipper, 2018; Cocks, Dipper, Middleton, & Morgan, 2011; Eggenberger et al., 2016 for effects of aphasia on speech–gesture integration in comprehension and production). The relative importance of cues may change dynamically based on the specific context and task at hand.

We may consider whether the experiments reported in the present study offer any evidence for dynamic weighting of cues. The speech–picture mismatch trials (the "no" trials) from Experiment 3 may offer some interesting insight: Although response times were longer overall in mismatch trials, there was no difference in accuracy between congruent and incongruent speech–gesture when speech did not match the picture. That is, participants were not distracted by the gesture matching the picture in the incongruent items when told to focus only on speech. Thus, the explicit task instruction to pay attention only to speech seems to have had an effect on the relative importance and weighting of gesture. Gesture is a cue that is not obligatory and thus not always an accompaniment to the speech signal. As such, it may play different roles, including a more clearly secondary role, depending on task demands. The equal weighting of speech and gesture cues found in Experiment 4 (replicating Kelly et al., 2010) may come about through a combination of task demands, that is, the explicit relevance of speech and gesture cues to the task (participants were told to pay attention to both), and language-specific properties (e.g., English being less focused on verbs in terms of information-packaging than Turkish).

For BSL, in Experiment 2, when both manual and mouthing cues were relevant, we saw that mouthing was a clearly secondary cue for all signers; the hands were a stronger and more reliable cue for meaning comprehension. For the group of (non-fluent) BSL learners, this was even more pronounced: Learners responded similarly to items in which the hands matched the picture, whether the mouthing was congruent or incongruent with the hands. Here as well, the weighting of the cues is affected by task demands. For learners, the demands on language comprehension are high—their comprehension of BSL is arguably more cognitively demanding and analysis of the phonetic forms on the hands requires more processing. In this context of lower accessibility of the primary signal, that is, of information from the hands, the mouthing contributes little (see Drijvers & Özyürek's, 2018, 2019 findings that gesture does not contribute to integration when information from speech is not accessible).

It is clear that further research is needed to investigate the factors that modulate the interaction between different channels of information to further our understanding of the

relative contribution and weight of different channels in different contexts and situations. Whether the type of information conveyed pertains to semantic or phonological information and the extent to which secondary cues obligatorily co-occur with the primary cue are further important factors, which are relevant to and reflected in the present findings. In studying effects on language learning, gestures have been shown to be beneficial to the meaning level but not to the phonetic or phonological level, regardless of whether the gestures themselves were semantically iconic of the word meaning (Kelly & Lee, 2012) or iconic of the sound-level information (Hirata & Kelly, 2010; Hirata et al., 2014, e.g., producing a beat-like gesture for a short vowel and a sweeping gesture for a long vowel). In looking at the contribution of information from the secondary channel in both language modalities, the present findings also support the idea that semantic (iconic) information is more important than phonological information. Effects of task demands notwithstanding, speech and gesture showed a bidirectional, mutual interaction in native speakers of English. The semantic association between gesture and (lexical affiliates in) speech may contribute to the importance of gestures in comprehension (as suggested also by the stronger weight of gesture in Turkish, Özer & Göksun, 2019).

In contrast, native/fluent signers of BSL did not exhibit an equal weighting of information from the hands and mouth when told to pay attention to both cues. As code-blended or borrowed elements, mouthings provide information related to the phonology of English (Vinson et al., 2010), usually semantically congruent with the co-occurring manual sign. It would be interesting to compare the role of mouth gestures in integration in comprehension studies, as mouth gestures may convey semantic information, adjectival (e.g., thick-thin; hard-soft), adverbial (e.g., intensely, casually), or verb-level (e.g., blowing). It may be that mouth gestures are integrated more strongly than mouthings, more comparably to co-speech gestures. In addition, some signs occur obligatorily (at least in citation form) with mouth gestures that are not considered to be semantic (e.g., "shh" occurring with the BSL sign NOT-YET, and other mouth gestures exhibiting echo phonology, Woll, 2009). These mouth gestures thus contribute information on a more phonological level, but may co-occur more obligatorily with signs than mouthings (though see Johnston et al., 2016).

Finally, for all of this, careful reflection on the term "integration" is warranted. In its use in the literature (the present paper not excepted), the term ranges from a strict sense of unification of two or more streams of information to a single/coherent representation (as, e.g., in vision, where integration is used to refer to the fusion of images from each eye) to a relatively loose sense of interaction or influence. The interaction between or influence of one channel on the other does not necessarily imply integration in the strict sense. By the same token, and as the analogy with vision may serve to illustrate, integration or unification in the strict sense does not entail that information from different sources is equally important or equally weighted. In vision, for example, stereoscopic fusion is arguably the most important cue for depth perception, but we can achieve depth perception through the use of other cues (e.g., texture or slant) when stereoscopic information is not available (e.g., if we are looking with only one eye).

5. Conclusion

Our findings from both language modalities are strong testimony to the multimodal nature of language. Studying both language modalities in parallel will continue to shed light on the nature of the integration and interaction between different cues and, in particular, the way in which different cues are dynamically weighted in context. In having cues from the face and mouth present, together with cues from the hands, we have contributed to the move toward more ecological validity, but we still know little about integration in face-to-face contexts. It remains to be seen how the many additional cues that are also available, such as other kinds of bodily movements, facial expressions and other acoustic information, are combined in real time when we communicate in the real world—treating all cues as context and understanding how they are used in regulating interaction.

Acknowledgments

This research was supported by UK Economic and Social Research Council grant ES/K001337/1 to D.V. and RES-620-28-6002 to the Deafness, Cognition and Language Research Centre (DCAL). We thank Neil Fox for being the model for the BSL videos and for conducting the BSL experiments.

Notes

1. We follow the convention of using upper case letters to refer to signs.
2. A closely related, more ideological, debate focuses on whether mouthings should be rejected (Hohenberger & Happ, 2001) or accepted (Ebbinghaus & Heßmann, 2001) as a constitutive part of sign language.
3. For the results and analyses presented in this section, we acknowledge that we need to be cautious about drawing strong conclusions concerning group comparisons because of the small group sizes.
4. In a preliminary analysis, we also compared objects and actions as semantic categories but found this factor did not affect responses. Most of the gestures accompanying object words (about 75%) were action-based and thus semantically similar to the gestures accompanying action words. The difference between the two categories is thus primarily at the level of the spoken word, and there is no evidence of differential processing of the two categories (as object vs. action) or as a result of the different semantic relationship between speech and gesture between the categories.

References

Adam, R. (2013). Unimodal bilingualism in the Deaf community: Language contact between two sign languages in Australia and the United Kingdom. Unpublished PhD thesis, University College London.

Azar, Z., Backus, A., & Özyürek, A. (2020). Language contact does not drive gesture transfer: Heritage speakers maintain language specific gesture patterns in each language. *Bilingualism: Language and Cognition*, *23*(2), 414–428.

Bank, R., Crasborn, O. A., & Van Hout, R. (2011). Variation in mouth actions with manual signs in Sign Language of the Netherlands (NGT). *Sign Language & Linguistics*, *14*(2), 248–270.

Bank, R., Crasborn, O., & van Hout, R. (2013). Alignment of two languages: The spreading of mouthings in Sign Language of the Netherlands. *International Journal of Bilingualism*, *19*(1), 40–55.

Bank, R., Crasborn, O., & van Hout, R. (2016). The prominence of spoken language elements in a sign language. *Linguistics*, *54*(6), 1281–1305.

Bates, D., Maechler, M., & Bolker, B. (2013). lme4: Linear mixed-effects models using S4 classes. R package version 0.999999-2.

Boyes-Braem, P. (2001). Functions of the mouthings in the signing of Deaf early and late learners of Swiss German Sign Language (DSGS). P. Boyes-Braem & R. Sutton-Spence *The hands are the head of the mouth: The mouth as articulator in sign languages* , (99–131). Hamburg: Signum.

Boyes-Braem, P., & Sutton-Spence, R. (Eds.). (2001). *The hands are the head of the mouth: The mouth as articulator in sign languages*. Hamburg: Signum.

Capek, C. M., Waters, D., Woll, B., MacSweeney, M., Brammer, M. J., McGuire, P. K., David, A. S., & Campbell, R. (2009). Hand and mouth: Cortical correlates of lexical processing in British Sign Language and speechreading English. *Journal of Cognitive Neuroscience*, *20*, 1220–1234.

Church, R. B., Kelly, S., & Holcombe, D. (2013). Temporal synchrony between speech, action and gesture during language production. *Language, Cognition and Neuroscience*, https://doi.org/10.1080/01690965.2013.857783.

Cocks, N., Byrne, S., Pritchard, M., Morgan, G., & Dipper, L. (2018). Integration of speech and gesture in aphasia. *International Journal of Language and Communication Disorders*, *53*(3), 584–591.

Cocks, N., Dipper, L., Middleton, R., & Morgan, G. (2011). What can iconic gestures tell us about the language system? A case of conduction aphasia. *International Journal of Language and Communication Disorders*, *46*(4), 423–436.

Crasborn, O. A., van der Kooij, E., Waters, D., Woll, B., & Mesch, J. (2008). Frequency distribution and spreading behavior of different types of mouth actions in three sign languages. *Sign Language & Linguistics*, *11*(1), 45–67.

Dimitrova, D., Chu, M., Wang, L., Özyürek, A., & Hagoort, P. (2016). Beat that word: How listeners integrate beat gesture and focus in multimodal speech discourse. *Journal of Cognitive Neuroscience*, *28*(9), 1255–1269.

Drijvers, L., & Ozyurek, A. (2017). Visual context enhanced: The joint contribution of iconic gestures and visible speech to degraded speech comprehension. *Journal of Speech, Language, and Hearing Research*, *60*, 212–222. https://doi.org/10.1044/2016_JSLHR-H-16-0101.

Drijvers, L., & Ozyurek, A. (2018). Native language status of the listener modulates the neural integration of speech and iconic gestures in clear and adverse listening conditions. *Brain and Language*, *177–178*, 7–17. https://doi.org/10.1016/j.bandl.2018.01.003.

Drijvers, L., & Özyürek, A. (2019). Non-native listeners benefit less from gestures and visible speech than native listeners during degraded speech comprehension. *Language and Speech*, https://doi.org/10.1177/0023830919831311.

Druks, J., & Masterson, J. (2000). *An object and action naming battery*. London: Psychology Press.

Ebbinghaus, H., & Heßmann, J. (2001). Sign language as multidimensional communication: Why manual signs, mouthings, and mouth gestures are three different things. In R. Sutton-Spence & P. Boyes-Braem (Eds.), *The hands are the head of the mouth: The mouth as articulator in sign languages* (pp. 133–151). Hamburg: Signum.

Eggenberger, N., Preisig, B. C., Schumacher, R., Hopfner, S., Vanbellingen, T., Nyffeler, T. , Gutbrod, K., Annoni, J.-M., Bohlhalter, S., Cazzoli, D., & Müri, R. M. (2016). Comprehension of co-speech gestures in aphasic patients: An eye movement study. *PLoS One*, *11*(1), e0146583.

Emmorey, K., Borinstein, H. B., & Thompson, R. (2005). In J. Cohen, K. T. McAlister, K. Rolstad, & J. MacSwan (Eds.), *Bimodal bilingualism: Code-blending between spoken English and American Sign Language* (pp. 663–673). Somerville, MA: Cascadilla Press.

Emmorey, K., Thompson, R., & Colvin, R. (2009). Eye gaze during comprehension of american sign language by native and beginning signers. *Journal of Deaf Studies and Deaf Education*, *14*(2), 237–243.

Ferrara, L. N., & Nilsson, A.-L. (2017). Describing spatial layouts as an M2 signed language learner. *Sign Language and Linguistics*, *20*(1), 1–26.

Fisher, C. G. (1968). Confusions among visually perceived consonants. *Journal of Speech and Hearing Research*, *11*(4), 796–804.

Furman, R., Küntay, A., & Özyürek, A. (2014). Early language-specificity of children's event encoding in speech and gesture: Evidence from caused motion in Turkish. *Language, Cognition and Neuroscience*, *29*, 620–634.

Giuostolisi, B., Mereghetti, E., & Cecchetto, C. (2017). Phonological blending or code mixing? Why mouthing is not a core component of sign language grammar. *Natural Language & Linguistic Theory*, *35* (2), 347–365.

Habets, B., Kita, S., Shao, Z., Özyürek, A., & Hagoort, P. (2011). The role of synchrony and ambiguity in speech-gesture integration during comprehension. *Journal of Cognitive Neuroscience*, *23*, 1845–1854.

He, Y., Gebhardt, H., Steines, M., Sammer, G., Kircher, T., Nagels, A., & Straube, B. (2015). The EEG and fMRI signatures of neural integration: An investigation of meaningful gestures and corresponding speech. *Neuropsychologia*, *72*, 27–42.

Hirata, Y., & Kelly, S. D. (2010). Effects of lips and hands on auditory learning of second-language speech sounds. *Journal of Speech, Language, and Hearing Research*, *53*(2), 298–310.

Hirata, Y., Kelly, S. D., Huang, J., & Manansala, M. (2014). Effects of hand gestures on auditory learning of second-language vowel length contrasts. *Journal of Speech, Language, and Hearing Research*, *57*, 2090–2101.

Hohenberger, A., & Happ, D. (2001). The linguistic primacy of signs and mouth gestures over mouthing: Evidence from language production in German Sign Language (DGS). In R. Sutton-Spence & P. Boyes-Braem (Eds.), *The hands are the head of the mouth: The mouth as articulator in sign languages* (pp. 153–189). Hamburg: Signum.

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

Holle, H., Obleser, J., Rueschemeyer, S. A., & Gunter, T. C. (2010). Integration of iconic gestures and speech in left superior temporal areas boosts speech comprehension under adverse listening conditions. *NeuroImage*, *49*(1), 875–884.

Holler, J., Kokal, I., Toni, I., Hagoort, P., Kelly, S. D., & Özyürek, A. (2015). Eye'm talking to you: Speakers' gaze direction modulates co-speech gesture processing in the right MTG. *Social Cognitive and Affective Neuroscience*, *10*, 255–261.

Holler, J., Schubotz, L., Kelly, S., Hagoort, P., Schuetze, M., & Özyürek, A. (2014). Social eye gaze modulates processing of speech and co-speech gesture. *Cognition*, *133*, 692–697.

Johnston, T., van Roekel, J., & Schembri, A. (2016). On the conventionalization of mouth actions in Australian Sign Language. *Language and Speech*, *59*(1), 3–42.

Kelly, S., Healey, M., Özyürek, A., & Holler, J. (2014). The processing of speech, gesture, and action during language comprehension. *Psychonomic Bulletin & Review*, *22*(2), 517–523.

Kelly, S. D., & Lee, A. L. (2012). When actions speak too much louder than words: Hand gestures disrupt word learning when phonetic demands are high. *Language and Cognitive Processes*, *27*(6), 793–807.

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

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

Kendon, A. (1972). Some relationships between body motion and speech. *Studies in Dyadic Communication*, *7*, 177–216.

Krahmer, E., & Swerts, M. (2004). More about brows. In Z. S. Ruttkay & C. Pelachaud (Eds.), *From brows to trust: Evaluating embodied conversational agents* (pp. 191–216). Dordrecht: Kluwer Academic 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*(3), 396–414.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, *82*(13), 1–26. https://doi.org/10.18637/jss.v082.i13.

Leonard, T., & Cummins, F. (2011). The temporal relation between beat gestures and speech. *Language and Cognitive Processes*, *26*(10), 1457–1471.

Liddell, S. K. (1980). *American sign language syntax*, The Hague: Mouton.

Loehr, D. (2007). Aspects of rhythm and gesture and speech. *Gesture*, *7*(2), 179–214.

Macedonia, M., Müller, K., & Friederici, A. D. (2011). The impact of iconic gestures on foreign language word learning and its neural substrate. *Human Brain Mapping*, *32*(6), 982–998.

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, *264*, 746–748.

Mohammed, T., Campbell, R., Macsweeney, M., Barry, F., & Coleman, M. (2006). Speechreading and its association with reading among deaf, hearing and dyslexic individuals. *Clinical Linguistics & Phonetics*, *20*(7–8), 621–630.

Mohr, S. (2012). The visual-gestural modality and beyond: Mouthings as a language contact phenomenon in Irish Sign Language. *Sign Language & Linguistics*, *15*(2), 185–211.

Morrel-Samuels, P., & Krauss, R. M. (1992). Word familiarity predicts temporal asynchrony of hand gestures and speech. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *18*(3), 615–622.

Muir, L. J., & Richardson, I. E. G. (2005). Perception of sign language and its application to visual communication for deaf people. *Journal of Deaf Studies and Deaf Education*, *10*(4), 390–401.

Nadolske, M. A., & Rosenstock, R. (2007). Occurrence of mouthings in American Sign Language: A preliminary study. In P. Perniss, R. Pfau, & M. Steinbach (Eds.), *Visible variation: Comparative studies on sign language structure* (pp. 35–61). Berlin: de Gruyter.

Obermeier, C., Dolk, T., & Gunter, T. C. (2012). The benefit of gestures during communication: Evidence from hearing and hearing-impaired individuals. *Cortex*, *48*, 857–870.

Obermeier, C., Kelly, S. D., & Gunter, T. C. (2015). A speaker's gesture style can affect language comprehension: ERP evidence from gesture-speech integration. *Social Cognitive and Affective Neuroscience*, *10*(9), 1236–1243.

Özer, D., & Göksun, T. (2019). Visual-spatial and verbal abilities differentially affect processing of gestural vs. spoken expressions. *Language, Cognition and Neuroscience*, 1–19. https://doi.org/10.1080/23273798.2019.1703016.

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

Pimperton, H., Ralph-Lewis, A., & MacSweeney, M. (2017). Speechreading in deaf adults with cochlear implants: Evidence for perceptual compensation. *Frontiers in Psychology*, *8*, 106.

Pyers, J. E., & Emmorey, K. (2008). The face of bimodal bilingualism: Grammatical markers in American Sign Language are produced when bilinguals speak to English monolinguals. *Psychological Science*, *19*(6), 531–535.

R Core Team (2013). *R: A language and environment for statistical computing*, Vienna, Austria: R Foundation for Statistical Computing. http://www.R-project.org/

Ross, L. A., Saint-Amour, D., Leavitt, V. M., Javitt, D. C., & Foxe, J. J. (2007). Do you see what I am saying? Exploring visual enhancement of speech comprehension in noisy environments. *Cerebral Cortex*, *17*(5), 1147–1153.

Sáfár, A., Meurant, L., Haesenne, T., Nauta, E., De Weerdt, D., & Ormel, E. (2015). Mutual intelligibility among the sign languages of Belgium and the Netherlands. *Linguistics*, *53*(2), 353–374.

Sandler, W. (1999). Cliticization and prosodic words in a sign language. In T. A. Hall & U. Kleinhenz (Eds.), *Studies on the phonological word (Current Studies in Linguistic Theory)* (pp. 223–254). Amsterdam: John Benjamins.

Schegloff, E. A. (1984). On some gestures' relation to talk. In J. M. Atkinson & J. Heritage (Eds.), *Structures of social action. Studies in conversation analysis* (pp. 266–296). Cambridge: Cambridge University Press.

Stamp, R. (2016). Do signers understand regional varieties of a sign language? A lexical recognition experiment. *The Journal of Deaf Studies and Deaf Education*, *21*(1), 83–93.

Stamp, R., Schembri, A., Evans, B., & Cormier, K. (2016). Regional sign language varieties in contact: Investigating patterns of accommodation. *Journal of Deaf Studies & Deaf Education*, *21*(1), 70–82.

Straube, B., Green, A., Bromberger, B., & Kircher, T. (2011). The differentiation of iconic and metaphoric gestures: Common and unique integration processes. *Human Brain Mapping*, *32*(4), 520–533.

Sueyoshi, A., & Hardison, D. M. (2005). The role of gestures and facial cues in second language listening comprehension. *Language Learning*, *55*(4), 661–699.

Sutton-Spence, R. (1999). The influence of English on British Sign Language. *International Journal of Bilingualism*, *3*(4), 363–394.

Sutton-Spence, R. (2007). Mouthings and simultaneity in British Sign Language. In M. Vermeerbergen, L. Leeson, & O. Crasborn (Eds.), *Simultaneity in signed languages: Form and function* (pp. 147–162). Amsterdam: Benjamins.

Sutton-Spence, R., & Day, L. (2001). Mouthings and mouth gestures in British Sign Language. In P. Boyes Braem & R. Sutton Spence (Eds.), *The hands are the head of the mouth: The mouth as articulator in sign languages* (pp. 69–85). Hamburg: Signum Verlag.

Sutton-Spence, R., & Woll, B. (1999). *The linguistics of British Sign Language: An introduction*. Cambridge: Cambridge University Press.

van de Sande, I., & Crasborn, O. (2009). Lexically bound mouth actions in Sign Language of the Netherlands: A comparison between different registers and age groups. *Linguistics in the Netherlands*, *26* (1), 78–90.

van Wassenhove, V., Grant, K. W., & Poeppel, D. (2005). Visual speech speeds up the neural processing of auditory speech. *Proceedings of the National Academy of Sciences*, *102*(4), 1181–1186.

Vinson, D., Thompson, R. L., Skinner, R., Fox, N., & Vigliocco, G. (2010). The hands and mouth do not always slip together in British Sign Language: Dissociating articulatory channels in the lexicon. *Psychological Science*, *21*(8), 1158–1167.

Vogt-Svendsen, M. (2001). A comparison of mouth gestures and mouthings in Norwegian Sign Language (NSL). In P. Boyes-Braem & R. Sutton-Spence (Eds.), *The hands are the head of the mouth: The mouth as articulator in sign languages* (pp. 9–40). Hamburg: Signum.

Willems, R. M., Özyürek, A., & Hagoort, P. (2007). When language meets action: The neural integration of gesture and speech. *Cerebral Cortex*, *17*(10), 2322–2333.

Woll, B. (2009). Do mouths sign? Do hands speak?: Echo phonology as a window on language genesis. In R. Botha & H. Swart (Eds.), *Language evolution: The view from restricted linguistic systems* (pp. 203–244). Utrecht: LOT Occasional Series.

Woll, B., & MacSweeney, M. (2016). Let's not forget the role of deafness in sign/speech bilingualism. *Bilingualism (Cambridge, England)*, *19*(2), 253–255.

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

Yang, J., Andric, M., & Mathew, M. M. (2015). The neural basis of hand gesture comprehension: A meta-analysis of functional magnetic resonance imaging studies. *Neuroscience & Biobehavioral Reviews*, *57*, 88–104.