Trust my gesture or my word:
How do listeners choose the information channel during communication?

Burcu Arslan[1]
Francis Ng[2]
Tilbe Göksun[1]
Nazbanou Nozari[2,3]

[1]Koç University, Istanbul, Turkey
[2]Carnegie Mellon University, Pittsburgh, PA, USA
[3]Center for the Neural Basis of Cognition (CNBC), Pittsburgh, PA, USA

Word count: 8379 (main text without abstract and references)

Abstract

Information can be conveyed via multiple channels such as verbal and gestural (visual) channels during communication. Sometimes the information from different channels do not match (e.g., saying *right* while pointing to the left). How do addressees choose which information to act upon in such cases? In two experiments, we investigated this issue by having participants follow instructions on how to move objects on the screen. Experiment 1 examined whether people's choice of channel can be altered by feedback favoring either the verbal or the gestural channel. In Experiment 2, there was no feedback and participants were free to choose either channel. We also assessed participants' verbal and visuospatial working memory capacities. Results showed that, when faced with contradicting information, there is a natural bias at the group level towards relying on the verbal channel, although this bias can be temporarily altered by probabilistic feedback. Moreover, when labels were shorter and of higher frequency, participants relied more on the verbal channel. In the absence of feedback, the capacity of individuals' visual, but not verbal, working memory determined reliance on one channel vs. the other. Collectively, these results show that information selection in communication is influenced by group-level biases, as well as properties of items and characteristics of individuals.

Keywords: *gesture comprehension, gesture-speech incongruency, working memory, communication, information processing*

Introduction

Language is multimodal. People often produce gestures when they speak. These *co-speech gestures* might semantically contribute to the message conveyed through the speech modality (McNeill, 1992). Listeners attend not only to speakers' speech but also to their gestures and integrate information coming from both modalities (e.g., Kelly et al., 2010). This integration process includes evaluating whether gestures' semantic properties are in line with the meaning extracted from speech (Wu & Coulson, 2005, 2007). When aligned with speech, observing gestures can enhance listeners' comprehension, encoding, and learning (for a review, see Özyürek, 2014). But what do listeners do when information from the verbal and gesture channels do not match? Imagine someone giving you directions and halfway through, they tell you to turn right but point to the left. Or imagine watching an instructional video on YouTube, but the verbal description does not match the object, or the location, pointed to. Would you follow the verbal command or the gesture? This paper examines this question from two angles: (a) Do people have a preferred channel for information processing, and can the choice of that channel be manipulated by feedback? and (b) are there individual differences driving the choice of preferred channel in communication?

Processing information from two channels

Although gesture and speech are often semantically related, they differ in terms of the channel through which they convey information (i.e., visual and auditory/verbal channels, respectively). These two sources of information share common neural resources (Özyürek et al., 2007; Skipper et al., 2009) and are often integrated during language production (Nozari et al., 2015) and comprehension (Kelly et al., 2010). The two channels can also compensate for one another. For instance, when the speech system is impaired, as in people with aphasia, the gesture system can play a compensatory role in communication by providing part of the information missing from the impaired speech (e.g., Akhavan et al., 2018).

Studies have shown that observing gestures can facilitate listeners' comprehension, encoding, and learning (see Özyürek, 2014, for a review). Both children and adults perform better in learning and comprehension tasks when they are presented with gestures, particularly gestures that represent objects, actions, or pointing at things, along with verbal input (Beattie & Shovelton, 1999; Cook et al., 2013; Kartalkanat & Göksun, 2020; Macoun & Sweller, 2016; McKern et al., 2021; Valenzeno et al., 2003). Similarly, Beattie and Shovelton (1999) suggested that observing iconic gestures for action and object representations significantly

improved people's narrative comprehension compared to when they were not presented with gestures. However, Dargue and Sweller (2018) demonstrated that iconic gestures' contribution to language comprehension is present only when those gestures are typical gestures that are semantically related to verbal content. In addition, Pi et al. (2017) showed that people who were presented with deictic gestures that involved pointing to content in a lecture were more likely to attend and comprehend the content than those who observed no gestures, indicating that deictic gestures might aid the comprehension process by directing attention.

Gestures can also thwart comprehension if they are irrelevant to speech (Kelly et al., 2010). When explicitly asked to detect a possible incongruency between speech and gestures in mismatch tasks, participants are usually slower and less accurate in their responses when presented with incongruent information from speech and gesture channels (Kelly, Heayley et al., 2015; Kelly, Özyürek, & Marquis, 2010; Özer & Göksun, 2020a). For example, Wu and Coulson (2014a, 2022) had participants watch video clips in which a speaker's speech and gesture were either semantically related or unrelated. Participants were then presented with photos and categorized them as being relevant or irrelevant to the video content (the picture-probe task). The authors demonstrated that participants' responses were slower and less accurate in this task when speakers' iconic gestures and speech were semantically unrelated than when they were related.

To summarize, these studies suggest that during communication, individuals process both speech and co-speech gestures and are sensitive to the discrepancy in the semantic content that the two channels signal. One question remains: in the face of discrepancy and in the absence of any explicit instructions on which channel to follow, how do listeners decide which information to rely on? Is there an inherent bias towards using one channel vs. another in humans and can such a bias be reversed by training? Or are there individual differences driving the channel choice? None of the above studies has investigated whether individuals' preference for verbal or gesture channels can be manipulated by feedback. Yet, individuals are likely to differ in how much they attend to gestures in discourse (Aldugom et al., 2021; Wu & Coulson, 2014a, 2022). Similar differences may underlie channel preference in communication.

Individual differences in gesture processing

Listeners differ in the extent to which they rely on speech or gesture modalities as a part of their communication style (see Özer & Göksun, 2020b, for a review). Research has suggested that working memory (WM) resources might be closely related to the processing of a multimodal message (Aldugom et al., 2021; Özer & Göksun, 2020a; Schubotz et al., 2021;

Wu & Coulson, 2014a,b; Wu et al., 2022). In the picture probe classification task described earlier, Wu and Coulson (2014a) showed that increasing working memory load, by asking people to memorize a sequence of spatial locations, decreased the benefit of congruent co-speech gestures in determining whether pictures were related or unrelated to the video clip. This finding demonstrated the importance of visuospatial WM resources for co-speech gesture processing. Importantly, taxing WM through a verbal task (memorizing a sequence of digits) did not have the same effect, pointing to the specificity of the WM resources involved in gesture processing. Moreover, Wu and Coulson (2014a) showed that individual differences in visuospatial WM, but not in verbal WM, predicted the benefit from adding related gestures to speech (see also Wu & Coulson, 2022). The authors suggested that the visuospatial WM memory might be responsible for processing gestures and information integration from that channel with speech (see also Momsen et al., 2021).

Studies have also indicated the role of verbal WM in language processing, particularly when the message conveyed via speech is not clear. For example, Schubotz et al. (2021) demonstrated that when auditory information was presented with background noise, addressees' verbal WM capacity was positively associated with benefiting from speakers' use of iconic gestures in a word recognition task. Verbal WM is also indicated when distractors appear in the auditory modality (Özer and Göksun. 2020a). In line with previous research (Kelly et al., 2010; Wu & Coulson, 2014), Özer and Göksun (2020a) found that comprehension was less effortful when gesture and speech provided congruent information. Critically, when participants were exposed to visual mismatch (e.g., presenting an illustration of an action along with a gesture that is semantically unrelated to that action), visuospatial WM scores positively predicted the speed and the accuracy of participants' responses. On the other hand, when participants were presented with verbal mismatch (e.g., presenting an action prime of a video along with auditory information that is semantically unrelated to that action), verbal WM score was positively associated with the speed and the accuracy of their responses.

A critical role for verbal WM in gesture processing is less obvious in other cases. For example, Momsen et al. (2020) used a dual task paradigm in which participants observed videos that presented gesture-speech pairs that were either congruent or incongruent with one another. While observing these videos, participants also held either 1 or 4 digits in working memory in "low" and "high" load conditions, respectively. EEG data were collected as participants watched the videos and listened to discourse. As expected, performance on the WM task was better in the low vs. high load condition but was unaffected by speech-gesture congruency. Thus, the behavioral data did not support a relationship between speech-gesture

integration and verbal WM. However, the authors found ERP differences in real-time word comprehension under higher WM load, pointing to a more subtle influence of verbal WM abilities in this task.

It is also important to note that previous research on individual differences has mainly focused on the comprehension of iconic gestures (e.g., Momsen et al., 2020; Özer & Göksun, 2020a; Schubotz et al., 2021; Wu & Coulson, 2014a, b; Wu et al., 2022). However, processing other types of gestures might also be affected by individual differences in cognitive resources (Aldugom et al., 2020). For instance, deictic gestures (i.e., pointing gestures) are closely associated with speech as they might form a bridge between human cognition and the physical world (Alibali & Nathan, 2012). Aldugom and colleagues (2020) focused on individual differences in benefiting from pointing gestures during math instruction and found that people with higher visuospatial, but not verbal, WM were better at learning the content when verbal information was accompanied by deictic gestures. However, when verbal information was presented without gestures, those with high verbal WM were better at learning the content.

To summarize, the literature reviewed above points to a clear role for WM in speech-gesture integration. It is easy to see why better visuospatial WM resources should help with faster and more accurate processing of gestural information; gestures are visual, and greater abilities in keeping visuospatial information in working memory can help with the integration of such information with verbal information. The opposite can also be true; higher verbal WM could, at least in theory, lead to easier processing of speech, which could in turn make the integration of speech with gestural information easier. The literature, however, seems to provide stronger support for the former. Verbal WM only helps under certain circumstances, for example, when speech is degraded or when irrelevant auditory information is presented to interfere with the processing of gestures. Importantly, all of these studies were concerned with *integration* of the information over the two channels, rather than *selecting* one channel of information over another, leaving open the question of which resource may be important for the latter purpose.

The current study

The present study examined individuals' choice of channel when gesture and speech provided incongruent information. In two experiments, we created interactive video clips, where participants first heard a set of instruction, accompanied by a hand pointing to objects and locations on the screen, and then carried out those instructions themselves in a different part of the screen. This task simulates many educational apps or interactive instructional clips

on YouTube or other online outlets, in that verbal instructions are accompanied by visual pointers, sometimes a hand and sometimes a cursor. Similarly, in our task, synchronously with verbal instructions, a hand figure visible to the participants pointed to the objects and the direction of the movement. The main manipulation was the congruency of verbal and visual information. In 1/3 of the trials, the verbal and gestural instructions were congruent. In the other 2/3, they were incongruent for either the object or the direction of the movement. In Experiment 1, we first provided probabilistic feedback after each trial favoring the verbal instructions or gestures in verbal and visual conditions, respectively, in a between-subject design. We then examined participants' performance after the removal of feedback. In Experiment 2, no feedback was provided. Experiment 1 allows us to test (a) if participants, as a group, have a default channel when they receive inconsistent information across channels, and (b) whether the choice of channel can be altered by feedback. If the latter, the experimental design also allows us to examine the persistence of such a change after the removal of feedback. Experiment 2 provides an opportunity to replicate any biases (i.e., natural tendencies in the absence of training) with a new and larger group of participants. Moreover, using a larger sample, we can investigate individual differences that may affect channel choice. Following previous studies, we used visuospatial and verbal WM capacity, but with a better-matched design. Recall that unlike past studies, the task goal here is not to integrate, but to select which channel of information to rely on in incongruent trials. The most straightforward prediction is that people with better verbal WM would default to the verbal channel and those with better visual WM, to the visual channel.

Experiment 1

The aim of Experiment 1 was to test whether channel choice was biased towards verbal or gesture instruction at the group level and whether statistical regularities in channel reliability could alter this bias. Participants played a game in which they moved objects on a screen in response to relatively complex verbal instructions and deictic gestures that were sometimes incongruent. The task was designed specifically to be demanding on working memory to mimic the real-life situations of maintaining and following complex multi-step instructions such as receiving directions. The experiment consisted of two blocks. In the first "training" block, participants received feedback. In the second "test" block, feedback was removed. Participants were randomly assigned to one of the two conditions: in the *verbal* condition, feedback in the form of *correct* or *incorrect* endorsed the choice of the verbal channel on 70% of the

incongruent trials. In the *visual* condition, the same type of feedback endorsed the choice of the gesture channel in the incongruent trials with a similar probability. Test blocks were identical across the two conditions.

If there is no systematic bias in channel choice, we expect equiprobable channel preference at the group level at the starting point of the training blocks in both conditions. Training would then determine whether an initial bias could be overcome. Finally, the test blocks would show the longevity of training effects, if any. In the absence of a strong inherent bias, long-lasting training effects could be expected. A strong inherent bias, on the other hand, could predict a quick loss of the training effects.

Two more manipulations were embedded in Experiment 1. The first one examined possible differences in resolving discrepancies between verbal and gesture instructions on objects vs. directions, as the latter is inherently more spatial and usually more error-prone in speech (e.g., left/right confusion is more common than cat/dog confusion; Corballis & Beale, 1976; Visser, 2016). The second one examined the influence of ease of lexical retrieval on channel choice. Words that are longer and lower in frequency are harder to retrieve (e.g., Balota & Chumbley, 1985; Marslen-Wilson & Tyler, 1980). Two sets of items, one animal with shorter high-frequency names and the other geometric shapes with longer low-frequency names, were administered in a counterbalanced order across the two blocks. If channel choice is determined by ease of processing, one would expect more reliance on the verbal channel for the shorter high-frequency items and a shift towards the gesture channel for the longer low-frequency items.

Methods

Participants

Since no prior studies had addressed the question posed in the current study, no effect size was available. Therefore, a predetermined sample of 32 native English-speaking participants (17 females; M_{age} = 19.94, *SD* = 1.84) was recruited from Prolific (https://prolific.co/) and the subject pool of Carnegie Mellon University, for cash and course credit, respectively. The study was approved by Carnegie Mellon's Institutional Review Board (IRB).

Materials and design

Two sets of stimuli were created with lexical frequencies adopted from SUBTLEX-US (Brysbaert & New, 2009). The easy set contained six animals (dog, elephant, lion, monkey, rabbit, and zebra) with an average log frequency of 3.00 (*SD* = 0.62), and an average length of

2 syllables (*SD* = 0.63). The difficult set contained six geometric shapes (cylinder, hexagon, oval, parallelogram, pyramid, and trapezoid) with an average log frequency of 1.31 (*SD* = 0.80; significantly lower frequency that the easy set, *t* = 0.002), and an average length of 3.17 syllables (*SD* = 0.98; significantly longer than the easy set; 0.034). Pictures corresponding to each item were 100x100 pixel black and white line drawings from Microsoft PowerPoints' icon collection (Figure 1).

Figure 1. Easy and difficult sets used in the experiment.

Using items in each set, 60 short 1280x720 pixels clips were created in Microsoft PowerPoint for each block. Each clip contained a slide, which was divided into an upper panel (instruction panel) extending approximately 200 pixels from the top, and a lower panel (action panel) (Figure 2). On each trial, all six objects appeared in the instruction and action panel in different configurations. A "direction panel" also appeared on the left side of the instruction panel, with four squares, corresponding to *above*, *below*, *left*, and *right*. The addition of this panel was necessary to disentangle directions from objects. For example, "above" could be demonstrated without pointing to an object and naturally evoking a joint representation, e.g., "above the rabbit." For each set, 60 audio files were recorded for delivering the verbal instructions. These instructions were spoken by a native English speaker and were always in the format of *A, B, and C move above/below/to the left of/to the right of D*, e.g., "The dog, the monkey, and the zebra move above the rabbit." There were 2.25 seconds in between each articulating two objects, or an object and a direction. Synchronously with the verbal instruction, the image of a hand pointed to each object. Direction was indicated using the direction panel (see Figure 2).

In each block, 20 of the 60 clips were *congruent* clips, meaning that the verbal and gestural instructions matched. The other 40 were *incongruent* clips. Of these, half of the clips contained a mismatch on an object (e.g., "The dog, the elephant, and the zebra move above the rabbit." pointing at the elephant instead of the monkey) and the other half a mismatch on

direction (e.g., pointing at right in the direction panel instead of above). We created a balanced design, such that each object and direction appeared equally often in the experiment and as the incongruent target. Incongruent pairing was also balanced. For example, each direction was paired equally often with each of the other three directions in the incongruent trials. This resulted in two blocks of 60 trials, one with animals and one with geometric shapes, administered to participants in a counterbalanced order.

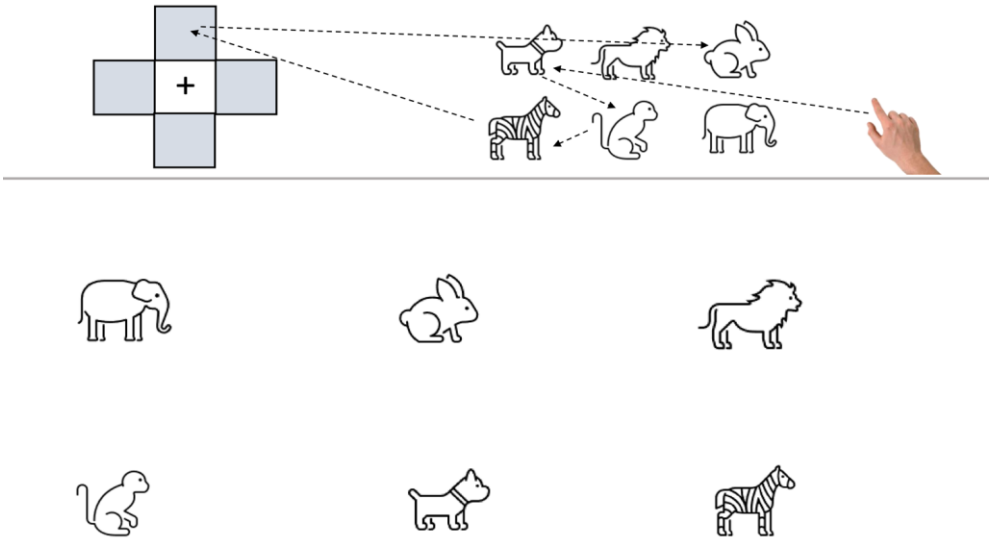


Figure 2. A still shot of an example video clip from the paradigm. The upper part is the instruction panel. The lower part, the action panel. A congruent verbal description for this trial would be "The dog, the monkey, and the zebra move above the rabbit.*"* An example of an incongruent trial would be "The dog, the *elephant,* and the zebra move above the rabbit." or "The dog, the monkey and the zebra move *to the right of* the rabbit."

Procedures

The experiment was developed in the jsPsych library (de Leeuw, 2015) and administered remotely with Jatos (Lange et al., 2015). through an internet browser on participants' personal computers. A sound test was then administered to ensure that participants had proper audio working on their personal devices. After passing the sound check, participants watched a video recording of instructions, which demonstrated how to carry out the task with two example trials. They then moved on to the experimental blocks. The two blocks of 60 trials, one with animals and one with geometric objects, were presented in a counterbalanced order. Before beginning each block, participants were introduced to the pictures in the set used in that block, along with their names. Then they were asked to type out each animal name to confirm they were familiar with those figures. After correctly identifying the objects for the first time,

two practice trials were given. The first practice trial had congruent instructions, while the second practice had incongruent instructions. If both practice trials were answered incorrectly, the participant would need to repeat the trial until getting it. For the incongruent trial, following either verbal or visual instructions would be accepted as a correct answer. Successfully completing both practice trials would cause participants to move to the training block.

On each trial of the training block, participants first watched the video clip, delivering simultaneous verbal and gestural instructions on how to rearrange objects in the Action Panel. They then followed the instructions by dragging and dropping objects in the specified locations using a mouse (see Figure 1). The task was self-paced. Participants clicked a *continue* button to signal that they were done. Then they received feedback as either "correct" or "incorrect". Feedback was determined only with regard to channel choice and was presented as "correct"/"incorrect". If participants chose the channel correctly but made other errors, the "correct" feedback would be displayed along with a warning prompting them to pay closer attention. Thus, participants had to probabilistically infer the preferred channel from binary feedback in an attention-demanding task.

The main manipulation was the probability of receiving the correct feedback as a function of instruction type (verbal vs. gesture) in the training block. Participants were randomly assigned to verbal and visual conditions. In the *verbal* condition, the feedback endorsed the verbal instructions on 70% of trials, and the gestural instructions on 30% of trials. In the *visual* condition, the reverse was true; the feedback endorsed the verbal instructions on 30% of trials, and the gestural instructions on 70% of trials.

After completing the training block, participants were given some time to rest before moving on the test block with the other set of objects. They typed out the name of each object in the new set and completed the block in a similar fashion to the first block, except with no feedback. The entire experiment took around 50 minutes.

The data and the analyses are available on the OSF: [https://osf.io/c7nz9/].

Analyses

Analyses were carried out using generalized linear mixed effect model (GLMM) in R (version 4.1.2, R Core Team, 2021), using lme4 library (Bates et al., 2015) and the lmerTest package (Kuznetsova et al., 2017). Choice of channel was predicted as a binary response (i.e., verbal or visual/gesture channel), with condition (i.e., verbal or visual training), block (i.e., test or training), and their interaction as fixed effects. Depending on specific questions, additional variables such as manipulation type (object vs. direction) or set (animal vs. geometric shapes) were added to the fixed effect structure. In addition, trial number within each block was

centered and entered into all models as a control variable to account for possible effects of item repetition or fatigue. To keep the random effect structure maximal and consistent across models (Barr et al., 2013). We included a structure that all models tolerated, consisting of the random intercept of subjects and items, as well as the random slope of block over subject. Categorical variables were coded as centered contrasts (-0.5, 0.5), and the numerical trial variable was centered.

Results

In total, there were 2560 incongruent trials across 32 participants. Out of these, 289 were discarded as participants followed neither the verbal nor the visual (gesture) instruction. After excluding these responses, we conducted the analysis with a total of 2271 incongruent trials. Figure 3 shows the results. Learning was calculated as the number of verbal choices divided by the sum of verbal and visual choices. For the visual condition, learning was calculated as the number of visual choices divided by the sum of verbal and visual choices. If there was no learning, one would expect choice to be at 50% (chance of randomly selecting one channel vs. another). As seen in the figure, participants in both conditions showed above-chance learning in the training block. Figure 4 unpacks this learning by dividing performance into quarters in each block.

The main model predicted channel choice as a function of condition, block, and the interaction between the two. Table 1 presents the results of this analysis. There was a significant effect of condition (β=1.60, z=3.90, p<.001), suggesting more verbal responses in the verbal than the visual condition (73% vs. 39%). There was also a significant effect of block β=-0.54, z=-2.14, p=.033), suggesting more verbal responses in the testing than the training block (58% vs. 52%). The effect of trial was marginal. There was also a marginal interaction between condition and block, pointing to the larger differences in the rate of verbal responses between the two conditions in the training vs. the test block.

Table 1
Model coefficients of a generalized linear mixed effect model predicting channel of choice. SE= Standard Error.

FIXED EFFECTS	Estimate	SE	z value	Pr(>|z|)
Intercept	-0.18	0.21	-0.89	0.38
Condition	1.60	0.41	3.90	<.001
Block	-0.54	0.25	-2.14	0.03
Trial	0.09	0.05	1.65	0.10
Condition*Block	0.85	0.50	1.70	0.09

RANDOM EFFECTS Intercepts	Variance	SD
Target Sentence intercept	0.02	0.14
Subject intercept	1.17	1.08
Slopes		
Block	1.51	1.23

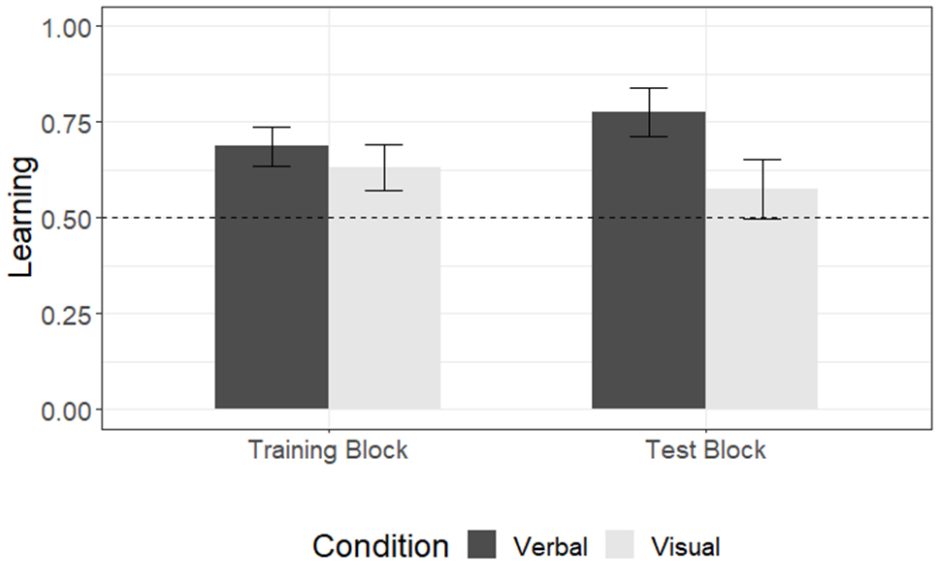


Figure 3. Learning in verbal and visual conditions in the training and test blocks. For the verbal condition, learning was calculated as the number of verbal choices divided by the sum of verbal and visual choices. For the visual condition, learning was calculated as the number of visual choices divided by the sum of verbal and visual choices.

The main analysis above showed a predominance of verbal choices especially in the verbal condition. In two follow-up tests, we examined the timeline and retention of learning in both groups (Figure 4). Specifically, we asked (1) How quickly is learning happening in verbal and visual conditions? (2) Do both groups retain this learning? To answer the first question, we examined learning at two time points, the end of the first and last block of training. To assess learning at the end of the first block, we compared performance on the first subblock of the training block (Train 1 in Figure 4) to chance (0.5) using a one sample *t*-test. Results, corrected for multiple comparisons, showed that the channel choice was significantly different than chance for the verbal condition ($t(15)=1.82$, $p=.04$), but not for the visual condition ($t(15)=1.31$, $p=.10$), suggesting a bias for the verbal channel established quickly in the first block. Note that this effect cannot be attributed to a greater tendency of participants in the

verbal group to choose the verbal channel. Examining the very first choice on an incongruent trial showed that 14 (88%) of the participants in the visual condition picked the verbal channel, while only 10 (63%) of participants in the verbal condition did so.

Next, we compared the performance in the fourth subblock of the training block to the first subblock of the training block (Train 4 to Train 1 in Figure 4) to assess the progress of learning in the two groups. Results of two-sample *t*-tests, corrected for multiple comparisons, indicated that there was learning in the visual condition ($t(15)=3.15$, $p=.014$) but not in the verbal group ($t(15)=.90$, $p=.720$). Together, these results imply different timelines of learning in the verbal and visual groups: in the verbal group, a preference towards picking the verbal channel was established quickly (by the end of the first subblock) and remained stable over the next three training blocks. In contrast, the visual group showed more gradual learning, remaining at chance at the end of the first subblock but showing robust learning by the end of training.

Next, we examined the retention of learning in both groups by comparing performance on the fourth subblock of the test block (Test 4 in Figure 4) to chance (0.5) to test whether the effects of the training persisted after the removal of feedback. The results of the one sample t-test showed that the choice of channel was different than chance for the verbal condition ($t(15)=4.92$, $p<.001$), but not for the visual condition ($t(15) = 1.14$, $p = .260$), suggesting that participants in the visual condition, as a group, reverted back to choosing the verbal channel.

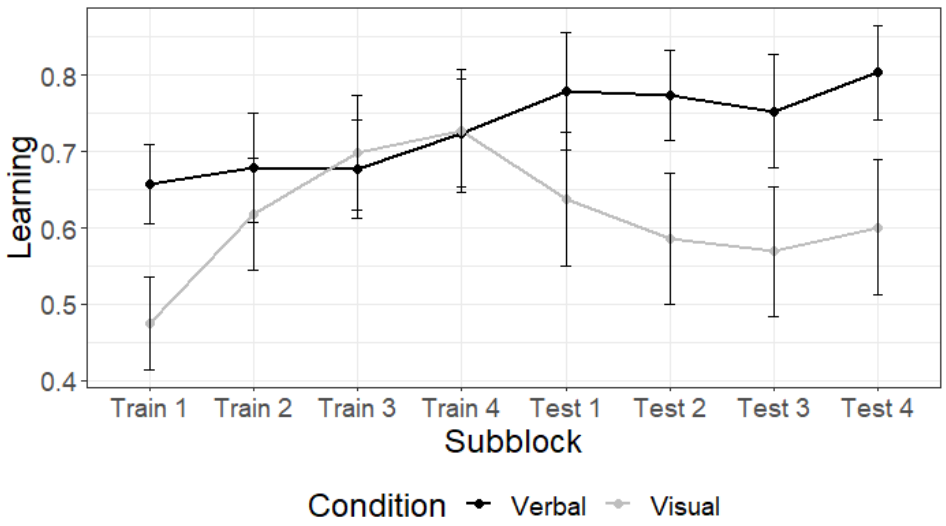


Figure 4. Breakdown of participants' learning by quarters in each block.

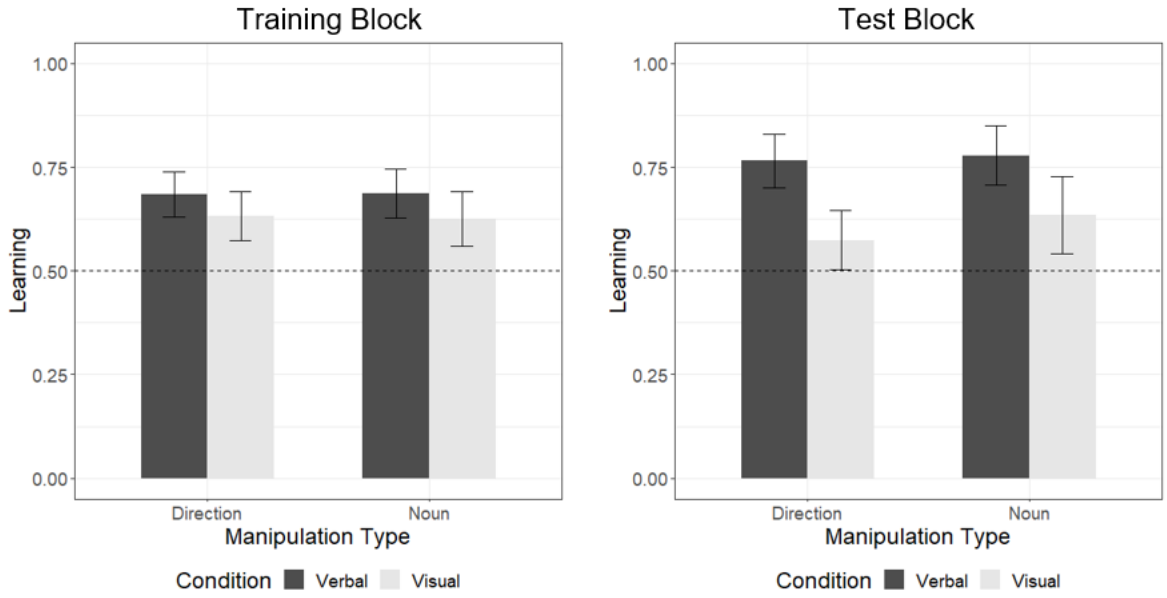
The effect of object vs. direction manipulation. In the next analysis, we explored whether the effect was different for objects and directions (Figure 5a, b). The model predicted the choice of channel as a function of condition, type of manipulation (i.e., direction or object), block, and their 2- and 3-way interaction between them. Similar to the main analysis, there were significant effects of condition (β=1.60, *SE*=0.41, p<.001), and block (β=-0.55, *SE*=0.25, p=.03) as well as a marginal interaction between condition and block. The effect of trial was also marginal. In contrast, neither the main effect of manipulation type, nor its interaction with condition, reached significance. Similarly, the 3-way interaction between condition, block, and manipulation type was not significant. In short, manipulating objects and directions had similar effects on performance[1] (Table 2).

Table 2
Model coefficients of a generalized linear mixed effect model predicting channel of choice. SE= Standard Error.

FIXED EFFECTS	Estimate	SE	z value	Pr(>|z|)
Intercept	-0.18	0.21	-0.88	0.38
Condition	1.60	0.41	3.89	<.001
Manipulation Type	-0.02	0.14	-0.15	0.88
Block	-0.55	0.25	-2.16	0.03
Trial	0.09	0.05	1.66	0.10
Condition * Manipulation Type	-0.07	0.27	-0.27	0.79
Condition * Block	0.85	0.50	1.69	0.09
Manipulation Type * Block	0.23	0.22	1.04	0.30
Condition * Manipulation Type * Block	0.44	0.43	1.03	0.30

RANDOM EFFECTS Intercepts	Variance	SD
Target Sentence intercept	0.02	0.13
Subject intercept	1.17	1.08
Slopes		
Block	1.51	1.23

[1] Some incongruencies are more common than others. For example, left and right are often more confused with one another than left/right with above/below. To ensure that the results were not different in subsets of data, we conducted three follow-up analyses, subsetting trials with (1) up/down or left/right incongruency (relatively common), (2) only left/right incongruency (most common), and (3) up/down exchanged with left/right (less common). All three analyses yielded the same effect of condition (i.e., learning) observed in the main analysis, showing that the main effect of interest was not an artifact of a subset of data.



(a) (b)

Figure 5. Learning in verbal and visual conditions for objects and directions in the training block (a) and the test block (b).

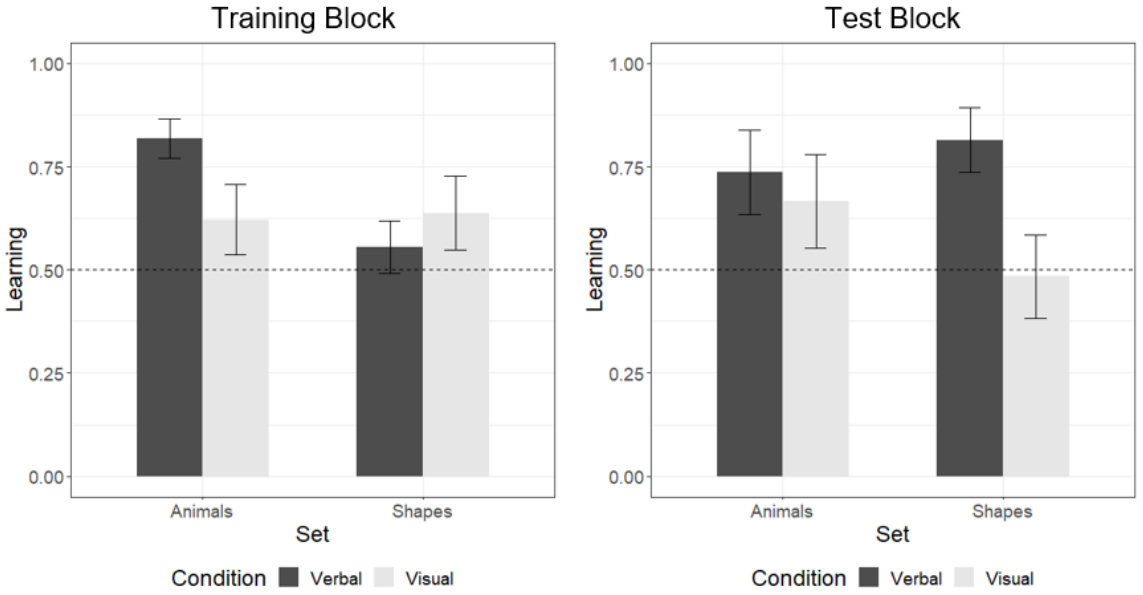
The effect of set. We also tested whether the effect was different for the easy vs. difficult sets (Figure 6a, b). The model predicted the choice of channel as a function of condition, set (i.e., animal or shape), block (i.e., training or test), and their 2- and 3-way interactions. Results, once again, showed a significant effect of condition (β=1.71, z=3.84, p<.001), block (β=-0.32, z=-2.94, p=.003), and a marginal interaction between the two (β=0.81, z=1.69, p=.09). We also observed a main effect of set (β=-0.51, z=-2.14, p=.033), such that there were overall more verbal responses for the animal than the geometric shapes set. In addition, there was a significant interaction between condition and set (β=-1.49, z=-2.03, p=.042), such that there were more verbal choices in the verbal compared to the visual condition for the animal (77% and 36%, respectively for the verbal and the visual conditions) vs. geometric shapes set (68% and %41, respectively for the verbal and the visual conditions). Other effects did not reach significance (Table 3).

Table 3
Model coefficients of a generalized linear mixed effect model predicting channel of choice. SE= Standard Error.

FIXED EFFECTS	Estimate	SE	z value	Pr(>|z|)

Intercept	-0.20	0.18	-1.08	0.28
Condition	1.62	0.37	4.41	<.001
Set	0.78	0.37	2.14	0.03
Block	-0.51	0.24	-2.14	0.03
Trial	0.09	0.05	1.64	0.10
Condition*Set	-1.49	0.73	-2.03	0.04
Condition*Block	0.81	0.48	1.69	0.09
Set*Block	-1.56	0.95	-1.69	0.10
Condition*Set*Block	1.18	1.91	0.62	0.54

RANDOM EFFECTS Intercepts	Variance	SD
Target Sentence intercept	0.02	0.14
Subject intercept	0.90	0.95
Slopes		
Block	1.32	1.15



(a) (b)

Figure 6. Learning in verbal and visual conditions for easy (animals) and difficult (geometric shapes) sets in (a) the training block and (b) the test block.

Finally, Figure 7 shows the individual differences in verbal bias in the combined data from the training and the test blocks. The verbal bias was computed as the difference between verbal and visual choices divided by the sum of those choices for each individual. As seen in

this figure, while most participants in each condition showed evidence of learning in their respective condition, there is quite a bit of variability in the data.

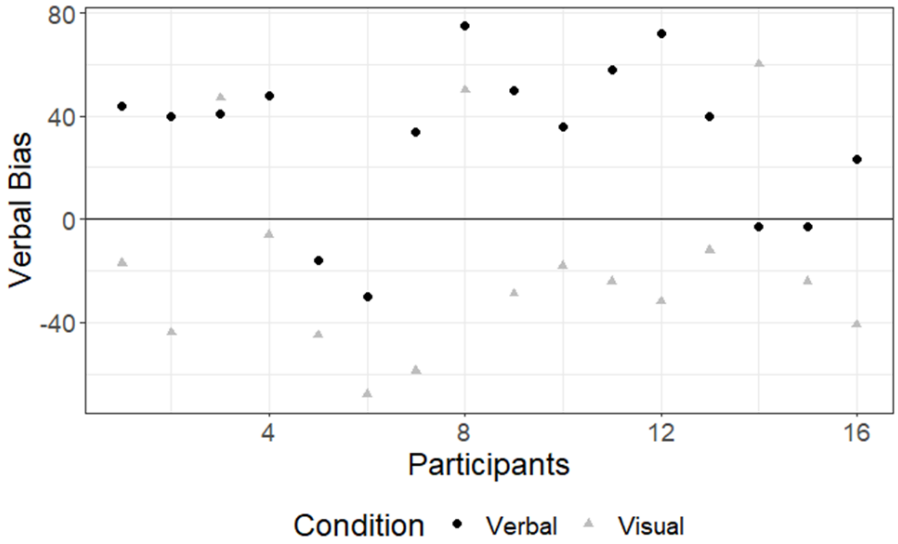


Figure 7. Verbal bias in individuals across the two conditions, collapsed over the two blocks. Verbal bias was calculated by dividing the difference between raw verbal and raw visual choices by the sum of raw verbal and raw visual choices.

Discussion

Results of Experiment 1 answered several questions. First, we observed a tendency towards choosing the verbal channel in participants in both verbal and visual conditions; this bias was quickly established in the verbal group during the first quarter of training and manifested as slower learning to rely on visual information in the visual group. Second, probabilistic feedback was successful in strengthening the bias in the verbal group and overturning it in the visual group. Third, learning was not long-lasting in the visual group. When feedback was removed, participants in this condition, as a group, reverted to relying on the verbal channel. Together, these findings suggest that when faced with conflicting information from verbal and gestural channels, listeners have a bias towards relying on the verbal channel, unless directed away from relying on this channel, in which case, they are capable of shifting to the gesture channel, at least temporarily.

While our manipulation of naming and pointing to objects was realistic, it is reasonable to object that the direction panel is not something people use in everyday life. To ensure that results were not tainted by the direction manipulation, we analyzed object and direction discrepancies separately. The pattern of results was identical, alleviating the concern that the

use of the direction panel may have distorted the results in any meaningful way. Finally, we examined the influence of indices of lexical retrieval, such as lexical frequency and length, on channel choice. We found that, as expected, the easier set evoked more verbal responses overall, and facilitated the reliance on the verbal channel in the verbal condition. These findings link the property of the materials, specifically the ease of lexical retrieval, to the use of the verbal channel.

All the effects reported above were at the group level. However, examining the performance of individual participants indicated much variability. Experiment 2 tested whether two factors that have been implicated in explaining individual differences in gesture processing, namely, visuospatial and verbal working memory capacities, could predict participants' choice of channel of information in the absence of external feedback.

Experiment 2

In Experiment 2, we used the same design of Experiment 1, except there was no feedback. Participants also completed variants of the Corsi Block Span task (Milner, 1971) and the forward Digit Span task (Wechsler, 2003), to obtain indices of their visuospatial and verbal working memory capacities, respectively. We then investigated whether scores on these tasks were correlated with individuals' likelihood of relying on the verbal channel (i.e., verbal bias).

Methods

Participants

No a priori effect size was available for the correlations investigated in the current study to formally calculate a necessary sample size. To estimate the sample size, we assumed a correlation of r = 0.35, which with α = 0.05 and a power of 0.90, returns a sample of 62 participants. We thus ran 64 native English-speaking participants (27 females; M_{age}=21.6, *SD*=2.12) recruited from Prolific (https://prolific.co/) and the subject pool of Carnegie Mellon University. Compensation and consenting were done in the same fashion as Experiment 1.

Materials and design

The same stimuli and design were used as Experiment 1. There were two blocks of 60 trials each, one with animals and one with geometric objects, with match (N=20) trials mismatch (N=40) present in each block. However, unlike Experiment 1, both blocks were test blocks. Participants did not receive any feedback. We also used the Corsi Block Span task and the Digit span task to measure individual differences in cognitive skills.

Corsi Block Span Task

In our adaptation of the Corsi Block Span Task (Milner, 1971; adapted for online use by Gibeau, 2021), 10 square blocks were displayed across a screen. In each of these trials, a sequence of blocks flashed on the screen for 400ms with 1400ms between each flash. After the sequence finished, a 500 hz beep played, prompting participants to click on the blocks in the order they had flashed on the screen. The sequence of blocks was the same for each participant for the trials. In the beginning of the task, following instructions, there were two practice trials with a sequence length of 3 blocks, where feedback was provided on performance. If a participant failed to get both practice trials correct, a third trial with feedback would play and repeat until completed correctly. Participants then started the experiment with two trials of sequence length of 3. If a participant completed one of the two trials correct, they were given two trials of sequence length 4, and so on. The experiment ended when participants either failed both trials for a given sequence length or finished the trials for sequence length 10. To calculate the score, we used a method, which would be sensitive to both block and trial number while at the same time keeping the scores within a reasonable range. The final Corsi score was obtained by dividing the number of correct trials for each participant by the total number of trials (16) and adding that decimal to participants' highest sequence length completed (block span). For example, if two participants both reached the span of 6, but one completed all the trials up to that level (i.e., 12) correctly and the other only one trial in each span up to that level (i.e., 6), the score for the first participant would be 6.75 and for the second participant 6.375.

Digit Span Task

The digit span task was adapted from Wechsler (2003) and was configured to closely matched the Corsi Block Span task in terms of procedures and number of levels. On each trial, a recording of a sequence of numbers was played, and participants typed the numbers in a box in the order they had heard them. The sequence of numbers was recorded using an editing program called *Descript* (https://www.descript.com/) with 500 ms intervals between each two digits. Procedures were similar to the Corsi Block Span task. The experiment began with two practice trials with feedback. If performed incorrectly, a third trial with feedback played and repeated until completed correctly. Participants were next given two trials per sequence length and needed to answer at least one of them correctly to proceed to the next sequence length. The experiment ended when the participant either failed both trials for a sequence length or finished the trials for sequence length 10. To calculate task score, the same method was used as in the Corsi task. We divided the number of correct trials for each participant by the total number of

trials (16) and added that decimal to participants' highest sequence length completed (digit span).

Procedure

All tasks were developed in JsPsych and administered remotely to participants through an internet browser on their personal computers using Jatos. After passing the sound check, participants first completed two blocks of the main task as described in Experiment 1, except that there was no feedback provided in either block. Upon completing the main task, participants completed the Corsi Block and the Digit Span tasks in fixed order. The experiment took roughly 55 minutes to complete.

Results

Nine participants were excluded because they did not follow the instructions (i.e., performed actions that were not aligned with either verbal or visual instructions on more than 2/3 of the trials). When we re-estimated power based on 55 participants, we found that it was still 0.85 for the hypothesized effect size and a type I error of < 0.05. Therefore, we proceeded and carried out the final analyses on 55 participants. Figure 8 shows the distribution of verbal bias, defined as the difference between verbal and visual choices divided by the sum of verbal and visual choices. As in Experiment 1, there was a bias towards using the verbal channel at the group level (67% of participants had a positive verbal bias). Also, there was considerable variability among participants in how biased they are towards using the verbal channel, motivating an analysis of individual differences.

The correlation between Corsi and digit span scores was small and non-significant ($r(55)=0.24$, $p=.075$), indicating low collinearity. As a further test of collinearity, we calculated the variance inflation factor (VIF) score. A VIF score lower than 2.50 is usually considered to pose no problem of collinearity (Shieh, 2011). Our model revealed a VIF score of 1.06, thus licensing us to enter the two WM scores as independent predictors in the linear regression analysis and examine their simultaneous influence on predicting the verbal bias. Results showed that the Corsi score was a significant predictor of verbal bias (β= -10.07, $t(54)$= -2.30, $p=.026$). However, the digit score did not significantly predict the verbal bias (β=1.99, $t(54)=0.45$, $p=.652$) (see Figures 9a, b).

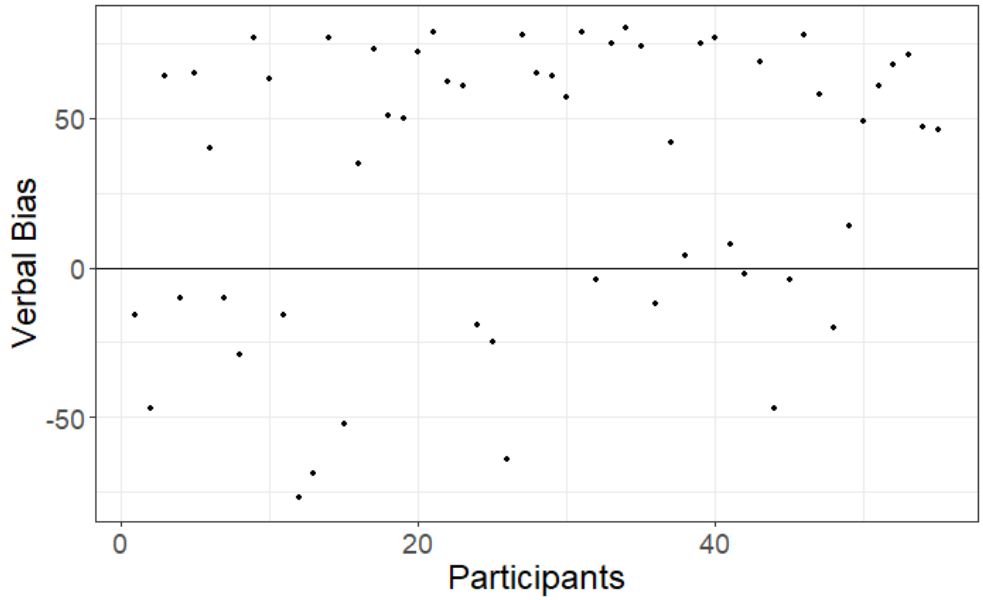
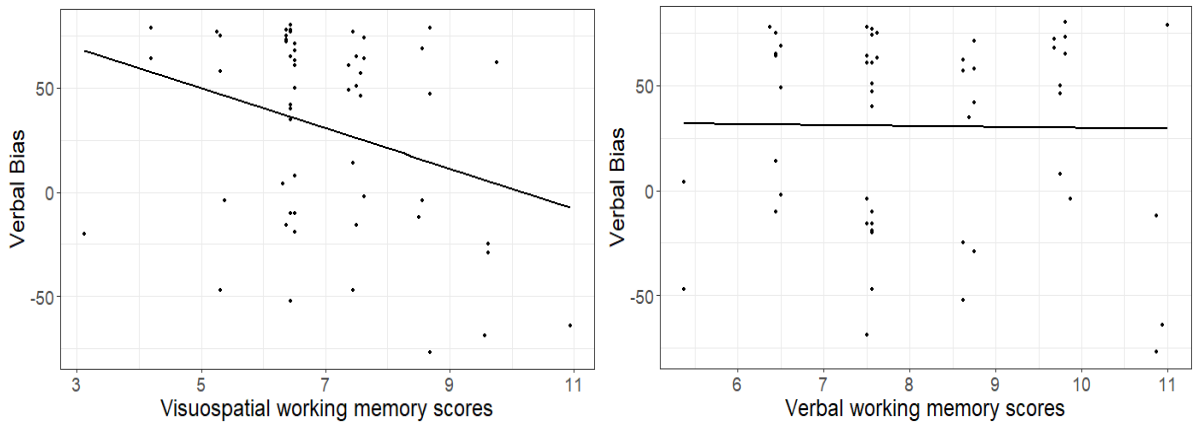


Figure 8. Individuals' tendency to choose the verbal channel.



(a) (b)

Figure 9. The distribution of visuospatial working memory scores (a) and verbal working memory scores (b) in relation to verbal bias. Pearson correlation between verbal bias and Corsi score was -0.30, (p=.027), and between verbal bias and digit span score -0.01 (p=.920).

Discussion

The aim of Experiment 2 was to investigate whether the variability in channel choice could be explained by people's verbal and visual working memory capacity. Replicating our findings from Experiment 1, Experiment 2 suggested a bias towards the verbal channel at the group level. Furthermore, the verbal WM capacity did not predict the bias in channel choice. Rather, it was the visual WM capacity that determined whether people deviated from this bias and relied more or less on the visual channel.

General discussion

Despite a large body of work on the *integration* of verbal and gestural information, to our knowledge, this is the first study investigating how participants *select* the information to rely on when not given explicit instructions. Our task was designed to mimic the complexity often encountered in everyday life. We often receive instructions through multiple channels, and sometimes these instructions clash, and we must choose which information to act upon. In two experiments, we first demonstrated that participants, as a group, show a bias towards using the verbal over gestural information. This bias, however, could be overcome by probabilistic feedback, although after removing feedback participants largely fell back on relying on verbal information. In addition, we found that the properties of the materials could affect the channel choice. When lexical processing was easier, participants relied more on verbal information. Despite these group-level tendencies, the data pointed to considerable variability among individuals, prompting the analysis of individual differences in Experiment 2. Experiment 2 replicated the variability, as well as the bias towards relying on the verbal channel, this time with no external feedback as participants chose freely which set of instructions to follow. We found that individual differences in visuospatial, but not verbal, WM were predictive of the choice of channel.

Our first finding of a bias towards the verbal channel is in line with studies showing that verbal labels can boost perception and memory (e.g., Lupyan et al., 2013; see Baddeley, 2003, for a review). When the task requires keeping multiple elements in mind to be executed in a certain order, verbal rehearsal is a known and effective strategy (e.g., Baddeley, 2000; Vallar & Baddeley, 1984). Results of our set difficulty manipulation further supported this idea. When lexical items were more difficult to process, because they were longer and less frequent (e.g., Balota & Chumbley, 1985; Marslen-Wilson & Tyler, 1980), participants relied less on the verbal channel. Additionally, learning in the verbal condition was better for the easier than the harder set. It is thus possible that simpler tasks reduce the bias observed towards relying on the verbal channel, but the conditions tested in this experiment better mimic real-life situations such as receiving multi-step instructions.

Our second finding, namely the flexibility of channel use as a function of probabilistic feedback, is also in line with studies showing humans' ability to learn under uncertain circumstances (e.g., Gluck & Bower, 1988; Poldrack et al., 1999). Participants in our study were remarkably fast in learning from probabilistic feedback. Almost halfway through the training block in the visual condition, participants' choice of the visual channel matched the

feedback probability of 70%, as has been observed in adult studies of statistical learning (e.g., Newport, 2020). This finding shows that, if desired, listeners can effectively shift their attention to gestural information even in the face of incongruent information from the verbal channel. Interestingly, however, this shift seems to be resource-demanding, as the removal of feedback in the test block caused the majority of participants to switch back to relying on the verbal channel. Collectively, these findings point to a flexible comprehension system that could use either verbal or visual information but has a preference to rely on verbal information, at least when the message's information content is high.

Even though directions are inherently more spatial than objects and depend on an abstract frame of reference system (Majid et al., 2004), we did not observe significant differences in participants' treatment of incongruencies on objects vs. directions. On the one hand, observing the same pattern for directions and objects is reassuring, given that our direction panel is not part of everyday communication. On the other hand, we acknowledge that more ecologically valid manipulations of directions may show that mismatches between verbal and gestural instructions on directions is even more impactful on participants' choices than mismatches on objects.

Finally, we observed great variability in channel choice in participants in Experiment 1. This finding is well aligned with prior studies showing individual differences in gesture processing (Aldugom et al., 2021; Momsen et al., 2021; Özer & Göksun, 2020a; Wu & Coulson, 2014a,b; Wu et al., 2022). While Experiment 1 showcased this variability in participants' learning from feedback, Experiment 2 demonstrated that the variability was not entirely learning-related; it was also present when listeners were left free to choose their preferred channel for information processing. Since the current design required maintaining sizeable chunks of information to act upon, a straightforward prediction was that participants with better verbal WM should rely more on verbal information, whereas participants with better visuospatial WM, should rely more on gestural information. To this end, we used the same verbal and visuospatial span tasks as previous studies. Methodologically, however, we improved upon prior designs by equating, to the extent possible, the design of Corsi Block Span and Digit Span tasks. Our adapted versions of the two tasks were matched on the procedures, as well as the number of levels participants could complete, thus providing more comparable measures than prior studies. In keeping with prior demonstrations, we found visuospatial and verbal WM resources to be largely separate (Cocchini et al., 2002), allowing us to investigate their separate influence on channel choice.

Our predictions were only partially confirmed. We found that participants' visuospatial WM capacities were predictive of their reliance on the gestural channel. This finding is in line with those of previous research, indicating the role of visuospatial WM in processing gestures (e.g., Özer & Göksun, 2020a; Momsen et al., 2021; Wu & Coulson, 2014a, 2022). As gestures are spatial representations (Alibali, 2005; Arslan & Göksun, 2021), understanding them requires spatial processing, which inevitably recruits visuospatial resources. Importantly, the current study extends the importance of visuospatial WM in speech-gesture integration (e.g., Aldugom et al., 2021; Wu & Coulson, 2014a, 2022) to the choice of gestural over verbal information to act upon.

In contrast, we did not find verbal WM capacity to be predictive of channel choice. This finding is in line with some prior studies. For example, Aldugom and colleagues (2020) found that visuospatial, but not verbal, WM scores were predictive of how much individuals benefitted from deictic gestures during math instructions. But what about studies that did find a relation between verbal WM and gesture processing (Kandana Arachchige et al., 2022; Momsen et al., 2020; Schubotz et al., 2021)? There are several points to note here. First, in some studies, e.g., Schubotz et al. (2021), speech was distorted, thus creating greater work for the verbal system to extract the signal, which may exaggerate the role of verbal WM in comprehension. Second, many studies include dual tasking (e.g., Kandana Arachchige et al., 2022; Momsen et al., 2020) which presents a special situation. Additionally, several such studies have failed to observe an overt effect of the verbal WM on gesture processing, sometimes despite a clear effect of visuospatial WM on performance. For example, Wu and Coulson (2014) did not find speech-gesture congruency benefits to be correlated with verbal WM (cf., Kandana Arachchige et al., 2022). In contrast, research from the same group reported ERP changes to comprehension as a function of WM load (Momsen et al., 2020). The authors attribute this discrepancy to the greater power of ERPs in picking up subtle differences in online processing. Critically, the nature of ERP differences between conditions was telling: under low WM load, speech-gesture congruency elicited a larger N400; under high WM load, a larger frontal positivity. The authors took these findings to mean automatic integration of speech and gestures for comprehension under normal circumstances and a role for verbal WM resources when this natural process fails. In other words, prior demonstrations of the role of verbal WM in gesture processing have generally pointed to a subtle role, often when processing demands were unusual.

Our results are compatible with this history. Specifically, the current findings extend those of Aldugom et al.'s (2020) by providing an explanation of why one type of WM matters

more than the other. One of our most robust findings was a verbal bias, meaning that participants had a natural tendency to use verbal, over gestural, information, perhaps because verbal communication is the dominant communication mode among adult speakers. Therefore, the majority of participants, regardless of their verbal WM status, defaulted to the more commonly used verbal channel for information processing. This default was, however, modulated by an ability to keep visuospatial information in mind. Those with lower visuospatial WM capacities relied even more strongly on the verbal channel, while those with high visuospatial WM capacities deviated from the group-level bias by using the gestural information more.

Limitations and future directions

At first glance, there might be concerns regarding the ecological validity of the task and the generalizability of the results. In real life, the hand performing the gestures is not disconnected from the speaker who produces the speech. It is then reasonable to wonder if these results apply to real-life situations. However, the aim here, unlike many prior studies, was not to test how listeners integrate speech and gesture in multimodal language processing (e.g., Mamus et al., 2023; Rasenberg et al., 2022). Rather, the question is about how deictic gestures are used to process information in the face of incongruency with verbal instructions. Note that, unlike iconic gestures, deictic gestures often point to information in the external world, with the purpose of guiding the listeners' attention to specific objects or parts of space. This can be accomplished by using fingers or any other object that the speaker chooses to use, such as pointers or cursors on a screen. In that sense, the setting of this study closely mimics what is found in online instructional videos and can thus be directly linked to such situations in real life. The findings can also be cautiously extended to situations in which an interlocutor provides conflicting speech-gesture information, and participants must choose one channel to act upon. This extension, however, has some limitations. For example, individuals are usually very good at adjusting their expectations based on their interlocutors and these adjustments are stable as long as the interlocutor is present (Holler & Bavelas, 2017). It is unclear whether the instructor's identity is equally strongly represented in instructional videos, like the paradigm used here. It thus remains possible that the dissipation of learning after removing feedback in the test blocks, observed here, would have a different timeline in live conversations. Moreover, real-life interlocutors may not produce as many incongruencies as tested in the current study. Future research could parametrically manipulate the degree of incongruency to examine the

minimum proportion of incongruent trials necessary for adjusting the reliance on one channel vs. another.

Another criticism is the redundancy of gesture-speech combinations, which could affect the results. Hostetter (2011) suggested that non-redundant gestures, particularly iconic gestures, might have larger benefits on communication than redundant gestures. However, we did not provide any information about redundancy. Gestures would have been redundant to speech if participants had been told that the verbal information was reliable. The critical point of our study was that participants did not know which information was more reliable. In addition, when it comes to redundancy, deictic gestures serve a different purpose than iconic gestures, as they help link the visual information to speech (Bangerter, 2004). Finally, although we uncovered certain properties of the materials (e.g., lexical indices) and participants (visual WM) to be predictive of channel choice, we are far from claiming that this set is exhaustive. Our claim is merely that both sets of factors matter. This opens up many possibilities for future research.

Conclusion

We found a group-level verbal bias in following instructions, when verbal and gestural information did not match. We also found that this bias was stronger when lexical processing was easier, but could nevertheless be altered by probabilistic feedback, at least temporarily. In the absence of feedback, the capacity of individuals' visual, but not verbal, working memory determined reliance on one channel vs. the other. Collectively, these results show that information selection in communication is influenced by group-level biases, as well as properties of items and characteristics of individuals.

Acknowledgements

We thank Demet Özer for her contribution to the initial stages of this study. This research was partially funded by the James S. McDonnell Foundation Scholar Award in Understanding Human Cognition #220020510 to TG.

References

Akhavan, N., Göksun, T., & Nozari, N. (2018). Integrity and function of gestures in aphasia. *Aphasiology*, *32*(11), 1310-1335. doi.org/10.1080/02687038.2017.1396573

Aldugom, M., Fenn, K., & Cook, S. W. (2020). Gesture during math instruction specifically

benefits learners with high visuospatial working memory capacity. *Cognitive Research: Principles and Implications*, *5*(1), 1-12. doi.org/10.1186/s41235-020-00215-8

Alibali, M. W. (2005). Gesture in spatial cognition: Expressing, communicating, and thinking about spatial information. *Spatial Cognition and Computation*, *5*(4), 307-331. doi.org/10.1207/s15427633scc0504_2

Alibali, M. W., & Nathan, M. J. (2012). Embodiment in mathematics teaching and learning: Evidence from learners' and teachers' gestures. *Journal of the Learning Sciences, 21*(2), 247-286. doi.org/10.1080/10508406.2011.611446

Arslan, B., & Göksun, T. (2021). Ageing, working memory, and mental imagery: Understanding gestural communication in younger and older adults. *Quarterly Journal of Experimental Psychology, 74*(1), 29-44. doi.org/10.1177/1747021820944696

Baddeley, A. (2000). The episodic buffer: a new component of working memory?. *Trends in Cognitive Sciences*, *4*(11), 417-423. doi.org/10.1016/S1364-6613(00)01538-2

Baddeley, A. (2003). Working memory and language: An overview. *Journal of Communication Disorders*, *36*(3), 189-208. doi.org/10.1016/S0021-9924(03)00019-4

Balota, D. A., & Chumbley, J. I. (1985). The locus of word-frequency effects in the pronunciation task: Lexical access and/or production?. *Journal of Memory and Language*, *24*(1), 89-106. doi.org/10.1016/0749-596X(85)90017-8

Bangerter, A. (2004). Using pointing and describing to achieve joint focus of attention in dialogue. *Psychological Science*, *15*(6), 415-419. doi: 10.1111/j.0956-7976.2004.00694.x

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language, 68*(3), 255-278. doi.org/10.1016/j.jml.2012.11.001

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1–48. doi.org/10.18637/jss.v067.i01

Beattie, G., & Shovelton, H. (1999). Do iconic hand gestures really contribute anything to the semantic information conveyed by speech? An experimental investigation. *Semiotica, 123*(1–2), 1–30. doi.org/10.1515/semi.1999.123.1-2.1

Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, *41*(4), 977-990. doi.org/10.3758/BRM.41.4.977

Cocchini, G., Logie, R. H., Sala, S. D., MacPherson, S. E., & Baddeley, A. D. (2002). Concurrent performance of two memory tasks: Evidence for domain-specific working memory systems. *Memory & Cognition*, *30*(7), 1086-1095. doi.org/10.3758/BF03194326

Cook, S. W., Duffy, R. G., & Fenn, K. M. (2013). Consolidation and transfer of learning after observing hand gesture. *Child Development, 84*, 1863–1871. doi.org/10.1111/cdev.12097

Corballis, M., & Beale, I. (1976). *Psychology of left and right*. Hillsdale, NJ: Erlbaum. doi.org/10.4324/9781003049029

Daneman, M., & Carpenter, P. A. (1980). Individual differences in working memory and reading. *Journal of Verbal Learning and Verbal Behavior, 19*(4), 450–466. doi.org/10.1016/S0022-5371(80)90312-6

Dargue, N., & Sweller, N. (2018). Not all gestures are created equal: The effects of typical and atypical iconic gestures on narrative comprehension. *Journal of Nonverbal Behavior, 42*, 327–345. doi.org/10.1007/s10919-018-0278-3

De Leeuw, J. R. (2015). jsPsych: A JavaScript library for creating behavioral experiments in a Web browser. *Behavior Research Methods*, *47*(1), 1-12. doi.org/10.3758/s13428-014-0458-y

Gibeau, R.-M. (2021). The Corsi Blocks Task: Variations and coding with jsPsych. *The Quantitative Methods for Psychology*, *17*(3), 299–311. doi:10.20982/tqmp.17.3.p299

Gluck, M. A., & Bower, G. H. (1988). From conditioning to category learning: an adaptive network model. *Journal of Experimental Psychology: General*, *117*(3), 227. doi.org/10.1037/0096-3445.117.3.227

Holler, J., & Bavelas, J. B. (2017). Multi-modal communication of common ground. In R. B. Church, M. W. Alibali, & S. D. Kelly (Eds.), *Why Gesture? How the Hands Function in Speaking, Thinking and Communicating* (pp. 213–240). Amsterdam: John Benjamins. doi.org/10.1075/gs.7.11hol

Hostetter, A. B. (2011). When do gestures communicate? A meta-analysis. *Psychological Bulletin*, *137*(2), 297. doi.org/10.1037/a0022128

Kandana Arachchige, K. G., Holle, H., Rossignol, M., Loureiro, I. S., & Lefebvre, L. (2022). High verbal working memory load impairs gesture-speech integration. *Gesture*. doi.org/10.1075/gest.20028.kan

Kartalkanat, H., & Göksun, T. (2020). The effects of observing different gestures during

storytelling on the recall of path and event information in 5-year-olds and adults. *Journal of Experimental Child Psychology, 189*, 104725. doi.org/10.1016/j.jecp.2019.104725

Kelly, S., Healey, M., Özyürek, A., & Holler, J. (2015). The processing of speech, gesture, and action during language comprehension. *Psychonomic Bulletin & Review*, *22*(2), 517-523. doi.org/10.3758/s13423-014-0681-7

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*(2), 260–267. doi.org/10.1177/0956797609357327

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. doi.org/10.18637/jss.v082.i13

Lange, K., Kühn, S., & Filevich, E. (2015). "Just Another Tool for Online Studies"(JATOS): An easy solution for setup and management of web servers supporting online studies. *PloS One*, *10*(6), e0130834. doi.org/10.1371/journal.pone.0130834

Lupyan, G., & Ward, E. J. (2013). Language can boost otherwise unseen objects into visual awareness. *Proceedings of the National Academy of Sciences, 110*(35), 14196-14201. doi.org/10.1073/pnas.1303312110

Macoun, A., & Sweller, N. (2016). Listening and watching: The effects of observing gesture on preschoolers' narrative comprehension. *Cognitive Development, 40*, 68-81. doi.org/10.1016/j.cogdev.2016.08.005

Majid, A., Bowerman, M., Kita, S., Haun, D. B., & Levinson, S. C. (2004). Can language restructure cognition? The case for space. *Trends in Cognitive Sciences*, *8*(3), 108-114. doi.org/10.1016/j.tics.2004.01.003

Mamus, E., Speed, L. J., Rissman, L., Majid, A., & Özyürek, A. (2023). Lack of visual experience affects multimodal language production: Evidence from congenitally blind and sighted people. *Cognitive Science*, *47*(1), e13228. doi.org/10.1111/cogs.13228

Marslen-Wilson, W., & Tyler, L. K. (1980). The temporal structure of spoken language understanding. *Cognition, 8*(1), 1-71. doi.org/10.1016/0010-0277(80)90015-3

McKern, N., Dargue, N., Sweller, N., Sekine, K., & Austin, E. (2021). Lending a hand to storytelling: Gesture's effects on narrative comprehension moderated by task difficulty and cognitive ability. *Quarterly Journal of Experimental Psychology, 74*(10), 1791-1805. doi.org/10.1177/17470218211024913

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

University of Chicago Press.

Milner, B. (1971). Interhemispheric differences in the localization of psychological processes in man. *British Medical Bulletin, 27*, 272–277. doi.org/10.1093/oxfordjournals.bmb.a070866

Momsen, J., Gordon, J., Wu, Y. C., & Coulson, S. (2020). Verbal working memory and co-speech gesture processing. *Brain and Cognition, 146*, 105640. doi.org/10.1016/j.bandc.2020.105640

Momsen, J., Gordon, J., Wu, Y. C., & Coulson, S. (2021). Event related spectral perturbations of gesture congruity: visuospatial resources are recruited for multimodal discourse comprehension. *Brain and Language, 216*, 104916. doi.org/10.1016/j.bandl.2021.104916

Newport, E. L. (2020). Children and adults as language learners: Rules, variation, and maturational change. *Topics in Cognitive Science, 12*(1), 153-169. doi.org/10.1111/tops.12416

Nozari, N., Göksun, T., Thompson-Schill, S. L., & Chatterjee, A. (2015). Phonological similarity affects production of gestures, even in the absence of overt speech. *Frontiers in Psychology, 6*, 1347. doi.org/10.3389/fpsyg.2015.01347

Özer, D., & Göksun, T. (2020a). Visual-spatial and verbal abilities differentially affect processing of gestural vs. spoken expressions. *Language, Cognition and Neuroscience, 35*(7), 896-914. doi.org/10.1080/23273798.2019.1703016

Özer, D., & Göksun, T. (2020b). Gesture use and processing: A review on individual differences in cognitive resources. *Frontiers in Psychology, 11*, 573555. doi.org/10.3389/fpsyg.2020.573555

Özyürek, A. (2014). Hearing and seeing meaning in speech and gesture: insights from brain and behaviour. *Philosophical Transactions of the Royal Society B: Biological Sciences, 369*(1651), 20130296. doi.org/10.1098/rstb.2013.0296

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. doi.org/10.1162/jocn.2007.19.4.605

Pi, Z., Hong, J., & Yang, J. (2017). Effects of the instructor's pointing gestures on learning performance in video lectures. *British Journal of Educational Technology, 48*, 1020 – 1029. doi.org/10.1111/bjet.12471

Poldrack, R. A., Prabhakaran, V., Seger, C. A., & Gabrieli, J. D. (1999). Striatal activation during acquisition of a cognitive skill. *Neuropsychology*, *13*(4), 564. doi.org/10.1037/0894-4105.13.4.564

R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Retrieved from https://www.R-project.org/.

Rasenberg, M., Pouw, W., Özyürek, A., & Dingemanse, M. (2022). The multimodal nature of communicative efficiency in social interaction. *Scientific Reports*, *12*(1), 19111. doi.org/10.1038/s41598-022-22883-w

Schubotz, L., Holler, J., Drijvers, L., & Özyürek, A. (2021). Aging and working memory modulate the ability to benefit from visible speech and iconic gestures during speech-in-noise comprehension. *Psychological Research, 85*(5), 1997-2011. doi.org/10.1007/s00426-020-01363-8

Shieh, G. (2011). Clarifying the role of mean centring in multicollinearity of interaction effects. *British Journal of Mathematical and Statistical Psychology*, *64*(3), 462-477. doi.org/10.1111/j.2044-8317.2010.02002.x

Skipper, J. I., Goldin-Meadow, S., Nusbaum, H. C., & Small, S. L. (2009). Gestures orchestrate brain networks for language understanding. *Current Biology, 19*(8), 661–667. doi.org/10.1016/j.cub.2009.02.051

Valenzeno, L., Alibali, M. W., & Klatzky, R. (2003). Teachers' gestures facilitate students' learning: A lesson in symmetry. *Contemporary Educational Psychology*, *28*(2), 187-204. doi.org/10.1016/S0361-476X(02)00007-3

Vallar, G., & Baddeley, A. D. (1984). Fractionation of working memory: Neuropsychological evidence for a phonological short-term store. *Journal of Verbal Learning and Verbal Behavior*, *23*(2), 151-161. doi.org/10.1016/S0022-5371(84)90104-X

Visser, L. (2016). *Left-right confusion: Is it caused by verbal labelling difficulty?* [Master's thesis, Utrecht University]. studenttheses.uu.nl/handle/20.500.12932/21786

Wechsler, D. (2003). WISC-IV technical and interpretive manual. San Antonio, TX: Psychological Corporation.

Wu, Y. C., & Coulson, S. (2005). Meaningful gestures: Electrophysiological indices of iconic gesture comprehension. *Psychophysiology, 42*(6), 654-667. doi.org/10.1111/j.1469-8986.2005.00356.x

Wu, Y. C., & Coulson, S. (2007). How iconic gestures enhance communication: An ERP study. *Brain and Language, 101*(3), 234-245. doi.org/10.1016/j.bandl.2006.12.003

Wu, Y. C., & Coulson, S. (2014a). Co-speech iconic gestures and visuo-spatial working

memory. *Acta Psychologica, 153*, 39-50. doi.org/10.1016/j.actpsy.2014.09.002

Wu, Y. C., & Coulson, S. (2014b). A psychometric measure of working memory capacity for configured body movement. *PloS One, 9*(1), e84834. doi.org/10.1371/journal.pone.0084834

Wu, Y. C., Müller, H. M., & Coulson, S. (2022). Visuospatial Working Memory and Understanding Co-Speech Iconic Gestures: Do Gestures Help to Paint a Mental Picture?. *Discourse Processes, 59*(4), 275-297. doi.org/10.1080/0163853X.2022.2028087