Look or Listen? Which Channel do People Rely on More?

Francis Ng

Carnegie Mellon University

Dietrich College Honors Thesis

Advisor: Nazbanou Nozari

Abstract

In our day to day lives, we are sometimes presented with information from two different channels of information, and sometimes this information can be incongruent. For example, if someone were to give directions and point left but say *right* instead. In situations such as these, which information channel do people rely on? To answer this question, we conducted two experiments. In Experiment 1, participants followed instructions from verbal and visual channels, that were sometimes incongruent, to move objects on their screens. In the first (training) block, probabilistic feedback either biased participants towards the verbal or the visual channel. The second (test) block contained no feedback. We found statistical learning of the bias in both verbal and visual conditions in the training block, although there was a stronger bias towards the verbal channel and much variability among participants. To investigate the origin of this variability, Experiment 2 removed feedback. Instead, participants completed two additional tasks: the Corsi Block and Digit Span tasks, which measure visual and verbal working memory, respectively. The results showed that people with higher visuospatial working memory showed less bias towards relying on the verbal channel. Collectively, these data suggest a bias towards prioritizing verbal information unless people have strong visuospatial working memory or the environment signals that the verbal channel is less reliable.

Introduction

When people think of communicating with another person, they associate it most strongly with verbal communication. However, human communication is multimodal: we receive various simultaneous cues via auditory and visual channels, including content words referring to specific entities, gestures indicating specific directions (Kita et al., 2007). Moreover, information is not always reliable. For example, people make errors in both speaking and gesturing (Akhavan et al., 2018).

Gestures that accompany speech might facilitate cognitive processes (Kita et al., 2017). When certain individuals are presented with more difficult tasks, iconic gesturing appears to have a beneficial effect on narrative comprehension (McKern et al., 2021). Additionally, it was found that people with higher visual working memory tend benefit from having gestures when given instructions (Aldugom et al., 2020).

There is variance in how gestures are produced and processed across situations and individuals. Speakers not only differ in how much they benefit from using gestures, but listeners also differ in how much they attend to a speaker's gestures (for a review, see Ozer & Göksun, 2020). People with higher spatial working memory were found to be better at processing co-speech gestures as they were more sensitive to speech-gesture incongruency (Ozer & Goksun, 2020). It would be beneficial to look at cognitive individual differences when studying co-speech gestures.

While there has been much research looking at incongruency in instructions from speech and co-gesture speech before, not much research has been done looking into which channel people tend to have preferences over. When presented with incongruent information through speech and co-speech gestures, which information channel do people tend to look more towards? This project aimed to answer this question. I manipulated the congruency of verbal and visual information in a communication paradigm and examine two main sources of variance that I hypothesized would affect the choice of channel of information: external (Experiment 1) and internal (Experiment 2). External refers to factors outside of a

listener's cognitive system such as whether the probability that information from a channel is reliable. Internal refers to factors within the listener's cognitive system, e.g., lexical retrieval and visuo-spatial abilities. In these two experiments, I looked at how cognitive individual differences may influence a person's channel bias. Collectively, the findings of these two experiments will contribute to our understanding of information processing in noisy environments in general, and in communication settings in specific. Would two people with very different visuospatial working process a noisy environment differently? Analysis of individual differences will be particularly helpful in understanding communication problems and using compensatory channels to alleviate them (e.g., Akhavan et al., 2018).

Experiment 1

The goal of this experiment was to test participants' sensitivity to the accuracy of verbal/visual information. Participants completed a task, in which they followed instructions such as "the rabbit, the lion, and the monkey, move above of the dog" (as seen in figure 1), and moved objects on their own screen accordingly. Instructions were delivered verbally, but a hand also pointed to the referenced objects and directions. On one third of the trials, the verbal and visual instructions were congruent. On the other 2/3 trials they were incongruent. For example, upon hearing the sentence above, the hand would point to the elephant when the word "rabbit" was spoken. The critical manipulation was a probabilistic bias induced either in favor of the verbal or the visual instructions on incongruent trials: half of the participants received feedback that was 70% in favor of verbal instructions, and the other half, feedback that was 70% in favor of visual instructions. We investigated whether participants successfully shifted their attention to the more reliable channel as a function of this bias, and whether they maintained that behavior after feedback has been removed.

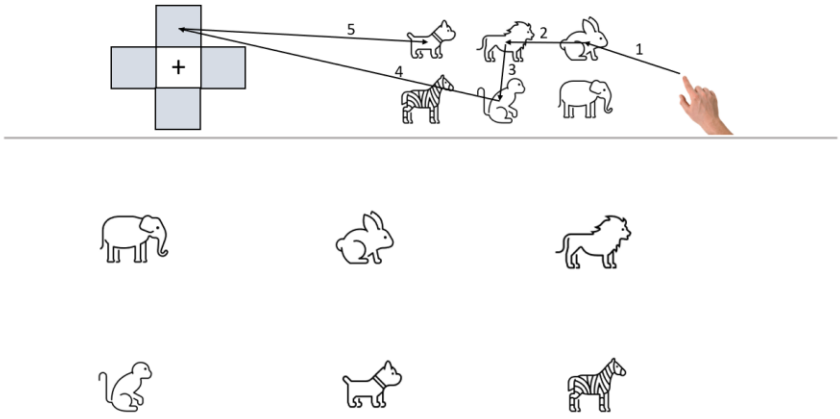

Figure 1. Example of a trial where numbers refer to the order the instructions name or gesture that object/direction

Additionally, we examined two other factors that could affect which channel is chosen: difficulty of lexical access (manipulated by having a low-frequency and a high-frequency item set), and manipulation of congruency on objects vs. directions. If the difficulty of lexical access is a determining factor in channel choice, we would predict that the more difficult set would lead to greater reliance on the visual channel.

Additionally, if processing directions (which contain spatial information) is more visually directed, we would expect greater reliance on the visual channel for direction incongruency than for object incongruency.

Methods

Participants A total of 32 participants were recruited (M_{age} = 19.9, *SD* = 1.84). 21 native English-speaking participants were recruited from Prolific (https://prolific.co/), and 11 native English-speaker participants were recruited from Carnegie Mellon University's participant pool using SONA. They were consented using an IRB protocol approved by the Carnegie Mellon University.

Stimuli and procedures. Two sets were created. The *easy* set contained six animals (dog, elephant, lion, monkey, rabbit, and zebra) with an average log frequency of 8.79 (SD = 1.32) adopted from the English Lexicon Project. The *difficult* set contained six geometrical shapes (cylinder, hexagon, oval, parallelogram, pyramid, and trapezoid) with an average log frequency of 6.28 (SD = 2.09). Pictures corresponding to each item were 100x100 pixel black and white line drawings from Microsoft PowerPoints' icon collection.

Using items in each set, 120 short 1280x720 pixels clips were created in PowerPoint. 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 1). The action panel was not used in the pilot study but was retained to keep the scene identical to the main experiments. For each trial, all six objects were arranged in different configurations for the instruction and action panel. These configurations remained consistent for every trial. A "direction panel" also always appeared on the left side of the instruction panel, with four squares, corresponding to "above", "below", "left" and "right". Sixty 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*, with A, B, C, and D being unique objects from a given set. 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.

There were a total of 60 trials per block. For each block, there were *congruent* (N = 20) and *incongruent* trials (N = 40). On the congruent trials, the verbal and visual instructions were the same. On the incongruent trials, they differed either for an object (N = 20) or a direction (N = 20). The same factors, e.g., equal assignment of incongruency to individual objects and directions, explained in detail in the pilot study were also controlled in this experiment.

Training and *test* blocks were created while the assignment of block type to set (i.e., animals vs. geometrical shapes) was counterbalanced across participants. The training block was always presented first and contained feedback after each trial. No feedback was given in the following non-training block. The trial structure was otherwise identical in the two blocks. On each trial, participants first watched the video clip, delivering simultaneous verbal and visual 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 mouse (see Figure 1). The task was self-paced. Participants clicked a *Continue* button to signal that they were done. In each trial of the training block, participants received feedback as either correct or incorrect in line with the condition they were in. No other information was given. In the test block, participants simply moved on to the next trial without feedback.

The main manipulation was the probability of receiving the *correct* feedback as a function of instruction type (verbal vs. gesture). In the "Verbal+ feedback" condition, the feedback endorsed the verbal instructions on 70% of trials, and the visual instructions on 30% of trials. In the "Visual+ feedback" condition, the reverse was true; the feedback endorsed the verbal instructions on 30% of trials, and the visual instructions on 70% of trials.

The experiment was developed in JsPsych and administered to participants through an internet browser on their personal computers. A basic demographics form was first given, collecting information such as age and sex. A sound test was then administered to ensure that participants had proper audio working on their personal devices. The sound test simply consisted of three trials, where a word was read aloud, and two pictures were displayed on the screen. The participant was required to click on the image that matched what was said. If this was done correctly all three times, they would proceed to the main experiment. If they failed, they would be given another set of three trials to attempt. If they failed to answer all the trials from the second set correctly, the experiment would end.

Participants were first shown a video of the experimenter demonstrating how to perform the task, and then completed two example trials. They then moved on to the experimental blocks. Before beginning each block, participants were shown the six animals/shapes along with their names and were asked to type out each animal/shape name to confirm they were familiar with each animal/shape. After correctly identifying the objects for the first time, two practice trial were given. The first practice trial had congruent instructions, while the second practice had incongruent instructions. If the practice trial was answered incorrectly, the participant would need to repeat the trial until getting it correct. For the incongruent trial, following either verbal or visual instructions would cause the participant to move to the training block.

After completing the training block, they would be given a small time of rest before moving on the testing block with the other set of objects. They would once again be asked to type out the name of each animal/shape for whichever one they will be doing for the testing block. The entire experiment took 50 minutes.

Analysis

To analyze the data, we ran generalized linear mixed-effect models. The first model ran contained condition, block, and the interaction between the two as fixed effects. The random effect structure included the random intercepts for both subjects and items, as well as random slopes of block over subjects. There was a significant effect on the condition (β=2.026, z=4.047 p<.001) and on the block (β=-0.5426, z=-2.128 p=.0333). No significant effect was found from the interaction between the condition and the block (β =.8556, z=1.695, p=.0900).

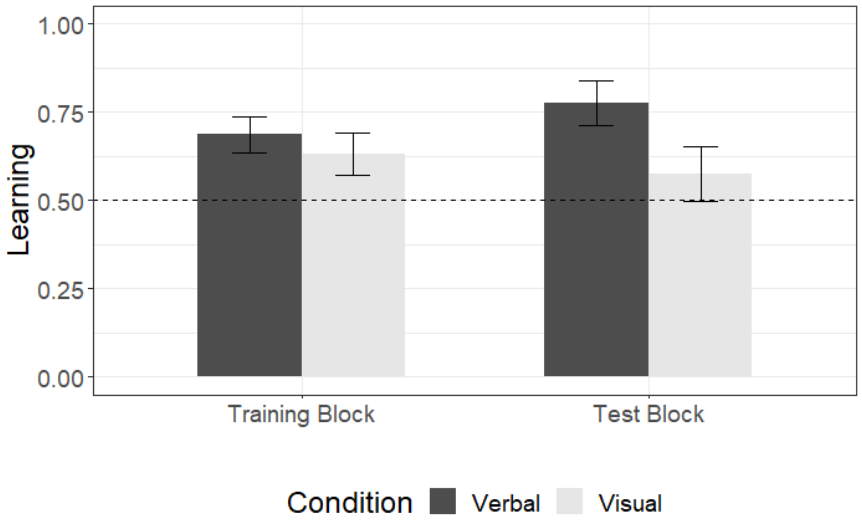

Figure 2. Choice of channel depending on the block. Learning is percentage of incongruent trials where the dominant channel was selected with standard error bars

We also looked at the learning present in each subblock of the experiment. From figure 3, we see people in the visual block started with no leaning towards any block, but over time learned to follow the visual channel instructions more. After the training block was over, it appears the learning gets weaker for the people in the visual channel. For people in the verbal channel, they immediately start learning, with learning score still remaining high for the testing block.

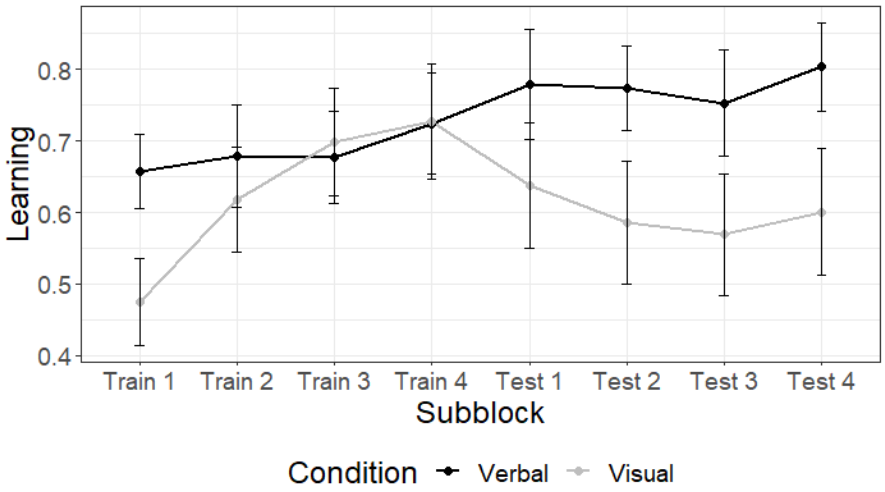

Figure 3. Choice in channel over the course of 8 subblocks with standard error bars.

We also wanted to see whether the effect was different for the incongruent instruction type (nouns or direction). We once again ran a generalized linear mixed-effect model containing condition, incongruent instruction type, and the interaction between the two. The random effect structure included the random intercepts for both subjects and items, as well as random slopes for incongruent instruction type over

subjects. We once again found a significant main effect from the condition (β=1.873, z=4.011, p<0.001), however there was no significant effect for the main effect of the incongruent instruction type (β=-.0362, -0.292, p=.8401) and the interaction (β=.2798, z=0.770, p=.4415).

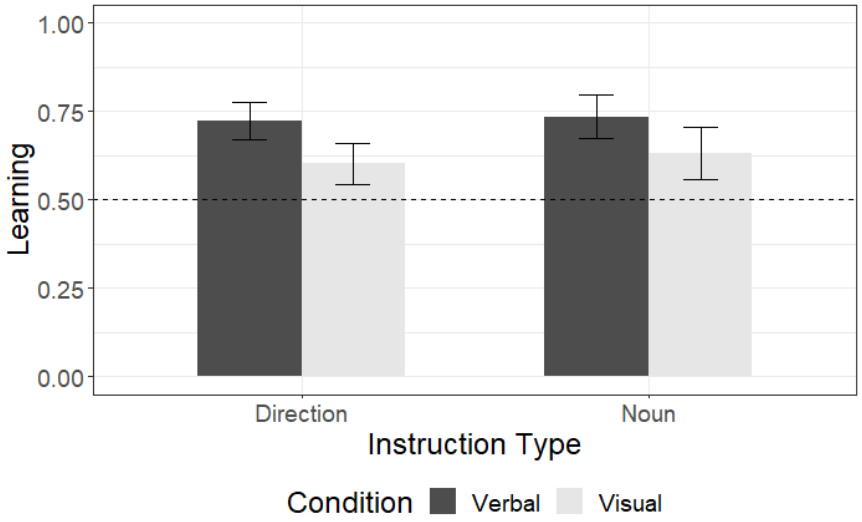

Figure 4. Choice of channel depending on which part of the instructions was manipulated with standard error bars.

Finally, we looked at whether difficulty of lexical access had an effect on the channel choice. We ran a generalized linear mixed-effect model containing condition, set, and the interaction between the two. The random effect structure included the random intercepts for both subjects and items, as well as random slopes of object set over subjects. The main effect of condition was found to be significant (β=1.983, z=4.078, p<0.001), while the main effect of the set was not significant (β=.0919, z=0.389, p=.6974). The interaction was found to be significant (β=-1.015, z=-2.143, p=.0321), prompting follow-up analysis on each set to see if there was learning in both.

Two generalized linear mixed-effect models were run, one subset on the animals set, and the other subset on the shapes subset. Both looked at the main effect of condition on channel choice, where the random effect structure included the random intercepts for subjects. For both the animals set (β=2.510, z=4.661, p<0.001) and the shape set (β=1.529, z=2.026, p=.0077), we found the main effect of condition was significant, with a larger coefficient for the animals set.

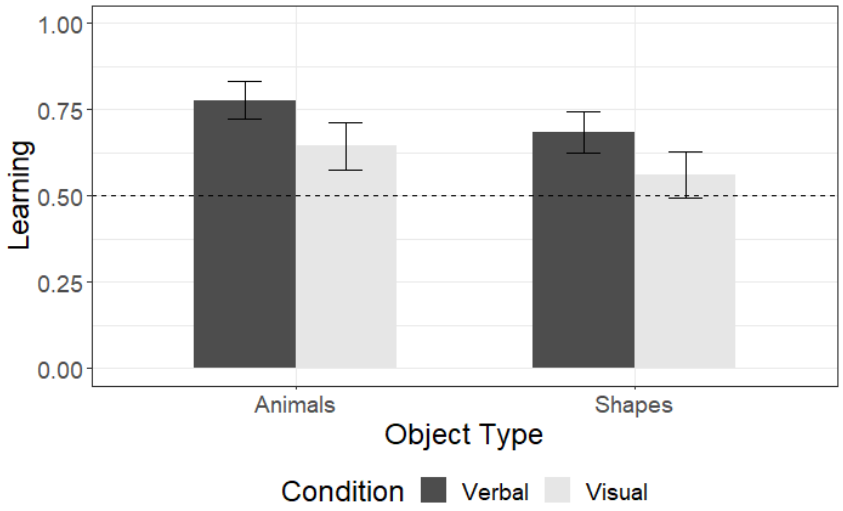

Figure 5. Choice of Channel depending on which object type was used with standard error bars.

Finally, we see in figure 6 that there is variability present among participants' verbal biases. While most participants in the verbal condition were above the 0 line, there were a few stray visual condition people who were also above the line, and vice versa for visual condition people below the line, showing a lot of variability of verbal scores among individuals.

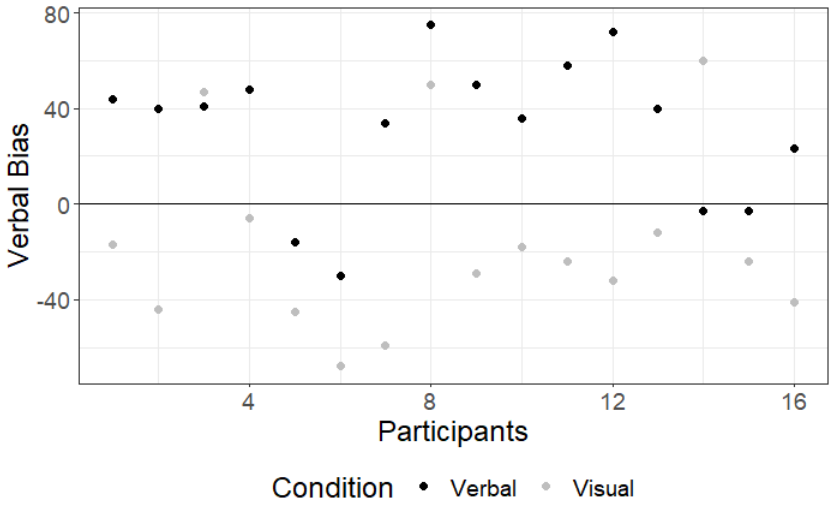

Figure 6. Verbal bias of the individual participants. Verbal bias is calculated by total trials where the participant selected verbal – total trials where the participant selected visual.

Discussion

This experiment aimed to find whether a person's choice of channel could be modulated by statistical learning. In figure 2, we saw that all the learning values were above 50%, showing that the choices that participants made were not random. Overall, we found that people will exhibit statistical learning.

From figure 3, there appeared to be learning present for both the verbal and visual condition of learning. In the training block, participants began to rely more on the assigned dominant channel. Participants in the visual condition usually began with more of a preference for the verbal channel, but they quickly picked up the statistical probability of the channels and began to pick the visual channel more often. However, looking at figure 3 these participants in the visual condition appeared to unlearn some of their bias towards the visual channel during the testing block when feedback was no longer provided and began showing a weaker bias towards the visual channel. Meanwhile, those in the verbal condition not only started with a slight bias towards the verbal channel, but their bias towards the verbal channel grew stronger throughout the experiment. This pattern is compatible with a natural bias towards relying on the verbal channel, as we typically associate communication with this information channel.

We also investigated whether statistical learning varied depending on whether the incongruency was from the nouns or the direction. When looking at the differences between noun and direction, there was no significant difference between the learning participants had for both. We commonly find ourselves identifying objects similarly to how they were identified in the experiment, however, our identification of directions is much more unnatural. There did not appear to be any significant interaction for the learning and the noun and direction instructions. This means that the way we administered the instructions did not impact the participants ability to learn the statistical probability.

Finally, we investigated whether lexical accessibility affected the channel choice. We found that people were able to learn the statistical probability regardless of whether they are viewing animals or geometric objects. While statistical learning did occur, it appeared to be weaker for the geometric objects than the animals. The difference in coefficients for the models we ran revealed while there was learning for both sets, there was a stronger affect for the animals compared to the shapes.

There appears to be clear evidence for statistical learning being present regardless of different types of information being manipulated (objects, directions). There also appears to be an inherent bias towards choosing the verbal channel, when examining individual participants in figure 6, we see there is variability with some participants showing little to no learning. This motivates us to look what internal factors could influence an individual's choices, which we will look at through examining individual differences of participants in the following experiment.

Experiment 2

In the experiment 2, we used the same design of the experiment 1, except there was no feedback provided. This experiment aimed to look at individual differences that might affect choosing between verbal and visual channels of information when faced with a set of incongruent directions. Participants completed the Corsi Block Span task, which measures visual working memory skills and the Digit Span task, which measures verbal working memory skills.

Methods

Participants A total of 64 participants (27 female; M_{age} = 21.6, SD = 2.12) were recruited. Nine native English-speaking participants were recruited from Carnegie Mellon University's participant pool using SONA, and 55 native English-speaking participants were recruited from Prolific (https://prolific.co/). They were all consented using the IRB protocol approved by the Carnegie Mellon University.

Stimuli and procedures. The same stimuli and trials were used as Experiment 1: a sound check and two practice trials were first done before proceeding to the main task. There were two blocks of 60 trials each, with *congruent* (N=20) trials and *incongruent* (N=40) trial present. Participants were once again counterbalanced for trial order and block order and watched a video clip with simultaneous verbal and visual instructions before dragging and dropping the object themselves. Participants did not receive any feedback after completing any of the trial.

Upon completing the main task, participants were introduced to the Corsi Block task. In this task, 10 square blocks are displayed across a screen (scaling with the size of the window) in set formation. In each of these trials, a sequence of blocks would flash on the screen for 400ms with 1400ms between each flash. After the sequence finishes, a 500hz beep would then play, prompting the participants to press the blocks in the order they flashed on the screen. The sequence of blocks was the same for each participant for the trials.

After reading the instructions, participants proceeded to complete 2 practice trials with a sequence length of 3 where feedback was provided on whether they memorized the sequence correctly. If they failed to get both practice trials correct, a third trial with feedback would play and repeat until done correctly. After practice, the experiment begins with two trials of sequence length of 3. If the participant gets one of the two trials correct, they will be given two trials of sequence length 4. The experiment ends when the participant either gets both trials for a sequence length incorrect or finishes the trials for sequence length 10.

The Digit Span task would be administered next. For each trial, a recording of a sequence of numbers being read aloud would play, and participants must type in a textbox what they believed the sequence to be. The sequence of numbers was recorded using a program called Descript to maintain consistency in the recording. There was a 500ms pause between each number being read aloud, where each sequence would be a number from 0-9 without any repeats. Similar to the Corsi Block task, each participant got the same sequence of numbers for each trial.

The procedure used for the Corsi Block test was mimicked to make the results easier to compare. The experiment begins with two practice trials that provided feedback on whether participants got the sequence correct. If not, a third trial with feedback would play and repeat until done correctly. Similar to the Corsi Block task, participants were 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 ends when the participant either gets both trials for a sequence length incorrect or finishes the trials for sequence length 10.

All three experiments were developed in JsPsych and administered to participants through an internet browser on their personal computers using Jatos. The code for the Corsi Block Task was found online and repurposed for running the experiment with the given guidelines (Gibeau, 2021). The experiment took roughly 55 minutes to complete.

Analysis

After running all the participants, some had to be excluded, as more than 1/3 of their responses were incorrect (incongruent trials where they selected neither the verbal nor visual channel). Nine participants were excluded due to this criterion, and we finished the analysis on the remaining 55 participants. Looking at figure 7, we find that most participants have a natural verbal bias.

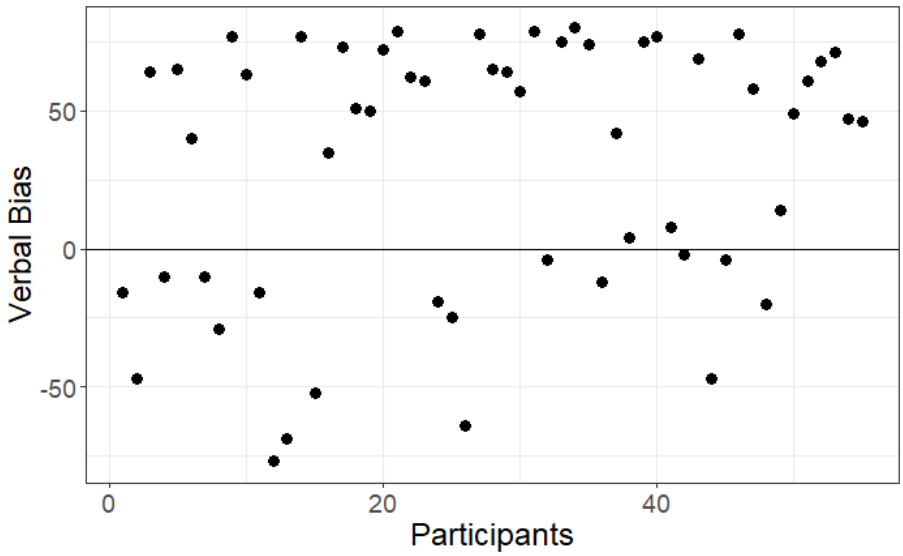

Figure 7. Individuals' tendency to choose the verbal channel with standard error bars.

Looking at scores for both the Corsi Block task and the Digit Span task were calculated by the highest sequence length the participant answered correctly added to the percentage of total trials answered correctly out of the 16 possible trials. In figure 8 and 9 below, we see that there is a negative slopped line for the visuospatial working memory in relation to verbal bias, while the line for the verbal working memory score had a flat line in relation to the verbal bias.

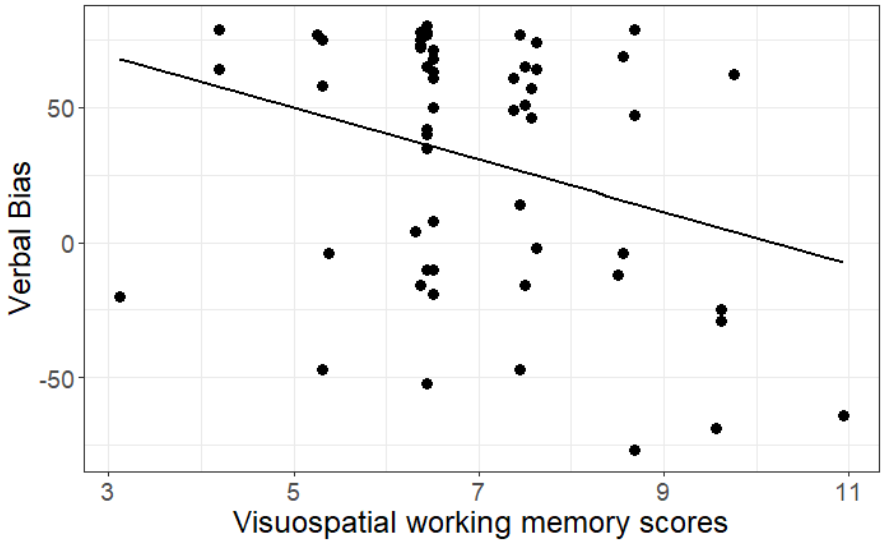

Figure 8. The distribution of visuospatial working memory scores with standard error bars.

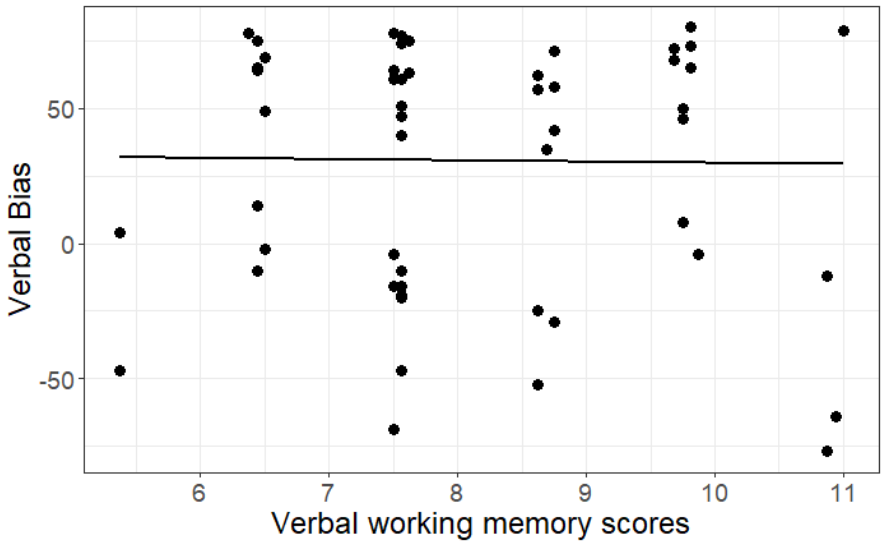

Figure 9. The distribution of verbal working memory scores with standard error bars.

Calculating the scores for both tasks gave us visuospatial working memory scores from the Corsi task (*M* = 7.00, *SD* = 1.46) and the verbal working memory scores from the digit span task (*M* = 8.15, *SD* = 1.45). There was no significant association between the two scores ($r(55)=.242$, $p=.075$). We conducted a linear regression analysis to predict the verbal bias with the Corsi score or the digit span score. The Corsi score was a significant predictor variable ($t(54) = -2.30$, $p = .026$), however the digit score was not a significant predictor variable ($t(54) = .45$, $p = .652$) for verbal bias.

Discussion

When feedback was never provided, we looked at whether individual cognitive differences were predictors in a person's intrinsic channel bias. We once again found there were an inherent bias towards the verbal channel. Verbal working memory score did not appear to be a significant predictor for their channel bias, but visuospatial working memory score was a significant predictor for channel bias. This is likely due to the inherent bias towards the verbal channel that we see no significance from the verbal working memory. However, people with a higher visuo-spatial working memory capacity seem to rely on the visual channel more, deviating from the inherent bias.

General Discussion

From the two experiments done, we found that people are faced with incongruent instructions, they will pick up the statistical bias when one channel is more reliable than the other. After feedback is removed, we found participants' whose dominant feedback channel was the visual channel started gravitating back towards the verbal channel, while participants' whose dominant feedback channel was the verbal channel continually stay strong in their natural bias. While no other manipulation studied had a significant main effect, we did find an interaction between the object set and condition. One possibility for this interaction is the lexical access of the geometric objects is much more difficult (Morsella & Krauss, 2004), thus participants have less resources for statistical learning from feedback. This could imply that when we were explaining more difficult concepts that tend to involve words with less lexical

access, using co-speech gestures may be helpful for visual learners, as they tend to pay more attention to their visual channels.

In the follow-up experiment, we looked at how individual differences affected channel decision and we found people with higher visuospatial working memory seem to rely on the visual channel more. Previous research has found that people with higher visuospatial working memory were better at processing co-speech gestures when facing incongruency between visual and verbal information, as these people had larger visuospatial cognitive resources (Ozer & Goksun, 2020). This aligns with what we found, as participants with smaller visuospatial working memory scores would have less cognitive resources available, making it simpler to rely more on the verbal channel for information.

Acknowledgments

Thanks to the Dietrich Seniors Honors Thesis Program and Joanne Ursenbach and Joseph Devine for giving me this opportunity. A huge thanks to Nikhil Lakhani, the lab manager of Nozari lab for assisting me with my numerous technical issues running my experiment. Another huge thanks to Burcu Arslan for assisting me with understanding the data and editing my thesis. Finally, and most importantly, a huge thanks to my advisor, Nazbanou Nozari, for teaching me so much about how to be a researcher and guiding me through the entire process.

References

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*. https://doi.org/10.1186/s41235-020-00215-8

Akhavan, N., Nozari, N., & Göksun, T. (2017). Expression of motion events in Farsi. Language, Cognition and Neuroscience, 32(6), 792-804.

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

Ezequiel Morsella, & Krauss, R. M. (2004). The Role of Gestures in Spatial Working Memory and Speech. *The American Journal of Psychology*, *117*(3), 411–424.

Kita, S., Özyürek, A., Allen, S., Brown, A., Furman, R., & Ishizuka, T. (2007). Relations between syntactic encoding and co-speech gestures: Implications for a model of speech and gesture production. Language and cognitive processes, 22(8), 1212-1236.

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

Özer, D., & Göksun, T. (2020). Gesture use and processing: A review on Individual Differences in Cognitive Resources. Frontiers in Psychology, 11:573555.

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

Zhang, X., Wu, Y. C., & Holt, L. L. (2021). The Learning Signal in Perceptual Tuning of Speech: Bottom Up Versus Top-Down Information. Cognitive Science, 45(3), e12947.