BRILL Multisensory Research (2023) DOI:10.1163/22134808-bja10094 brill.com/msr

Metacognition and Causal Inference in Audiovisual Speech

Faith Kimmet, **Samantha Pedersen**, **Victoria Cardenas**, **Camila Rubiera**, **Grey Johnson**, **Addison Sans**, **Matthew Baldwin and Brian Odegaard** *

Department of Psychology, University of Florida, Gainesville, FL 32603, USA
*Corresponding author; e-mail: bodegaard@ufl.edu
ORCID iDs: Baldwin: 0000-0002-7355-478X; Odegaard: 0000-0002-5459-1884

Received 6 April 2022; accepted 10 February 2023; published online 23 February 2023

Abstract

In multisensory environments, our brains perform causal inference to estimate which sources produce specific sensory signals. Decades of research have revealed the dynamics which underlie this process of causal inference for multisensory (audiovisual) signals, including how temporal, spatial, and semantic relationships between stimuli influence the brain's decision about whether to integrate or segregate. However, presently, very little is known about the relationship between *metacognition* and multisensory integration, and the characteristics of perceptual confidence for audiovisual signals. In this investigation, we ask two questions about the relationship between metacognition and multisensory causal inference: are observers' confidence ratings for judgments about Congruent, McGurk, and Rarely Integrated speech similar, or different? And do confidence judgments distinguish between these three scenarios when the perceived syllable is identical? To answer these questions, 92 online participants completed experiments where on each trial, participants reported which syllable they perceived, and rated confidence in their judgment. Results from Experiment 1 showed that confidence ratings were quite similar across Congruent speech, McGurk speech, and Rarely Integrated speech. In Experiment 2, when the perceived syllable for congruent and McGurk videos was matched, confidence scores were higher for congruent stimuli compared to McGurk stimuli. In Experiment 3, when the perceived syllable was matched between McGurk and Rarely Integrated stimuli, confidence judgments were similar between the two conditions. Together, these results provide evidence of the capacities and limitations of metacognition's ability to distinguish between different sources of multisensory information.

Keywords

metacognition, causal inference, audiovisual speech, multisensory, McGurk

DOI:10.1163/22134808-bja10094

1. Introduction

Our brains quickly and effortlessly integrate incongruent audiovisual sensory signals to produce a coherent perception of the world. For example, the presentation of incongruent numbers of brief flashes and beeps often yields illusory perception of phantom flashes (Shams *et al.*, 2000). Simultaneous, spatially discrepant audiovisual signals can cause the auditory component to be mislocalized close to where the visual component occurred (Pick *et al.*, 1969; Welch and Warren, 1980). And incongruent audiovisual speech syllables (e.g., auditory “Ba” and visual “Ga”) can produce perception of a unique, third syllable (e.g., “Da”) (McGurk and MacDonald, 1976). These examples of multisensory integration are often referred to as “illusions” in the scientific literature (Stevenson *et al.*, 2012) and reflect a general principle: when sensory cues are in conflict with one another, our brains have specific mechanisms which reconcile differences to produce unified, integrated perception (Ernst and Banks, 2002; Knill and Richards, 1996; Körding *et al.*, 2007).

While behavioral and computational studies of these multisensory illusions abound, little is known about the relationship between metacognition and multisensory integration. Metacognition can be generally defined as “thinking about thinking” (Flavell, 1979); in perceptual paradigms, metacognition can be measured by obtaining confidence ratings in perceptual decisions (Fleming and Lau, 2014). While the study of visual metacognition is well-established (Rahnev *et al.*, 2022), research on the relationship between multisensory integration and metacognition is limited to only a few studies (White *et al.*, 2014), and how metacognitive judgments interact with the process of “causal inference” that influences sensory integration remains almost entirely unexplored (Deroy *et al.*, 2016; Shams and Beierholm, 2010, 2021). In this investigation, we ask the following questions: does the average level of confidence differ for sensory information that is integrated from discrepant sources, segregated from multiple sources, or arises from only a single source? And more specifically, if the reported percept across different scenarios is the same, does the average level of confidence distinguish between congruent speech, integrated speech, and segregated speech?

Answering these questions is critical to better understand the relationship between metacognition and the hierarchical computations that form the basis of all sensory experiences, including multisensory integration (Körding *et al.*, 2007; McGovern *et al.*, 2016; Rohe and Noppeney, 2015; Rohe *et al.*, 2019). Currently, multisensory integration is thought to reflect the principles of Bayesian causal inference, as audiovisual cues are integrated or segregated based on a combination of noisy sensory encoding and priors that govern perception of a common cause (Magnotti *et al.*, 2018; Odegaard *et al.*, 2015; Rohe and Noppeney, 2015). This process of inference has been accounted for by Bayesian models (Körding *et al.*, 2007; Magnotti and Beauchamp, 2017; Magnotti *et al.*, 2013, 2018; Shams and Beierholm, 2010) which assume that observers have access to noisy sensory representations, but must infer the underlying causes of information which *generated* the signals. In a McGurk speech perception task, the number of causes can either be one ($C = 1$) if the stimuli are integrated, or two ($C = 2$) if the stimuli are segregated.

It is currently unknown whether our metacognitive awareness of this process is limited to the integrated end product of the sensory inference, or whether it can distinguish between congruent, incongruent, and illusory sensory content produced by various causal structures, especially when the final percept is the same (Fig. 1). Previous research provides preliminary evidence that metacognition may be able to distinguish among different causal scenarios (White *et al.*, 2014), but supporting data is sparse, and more research is needed.

In the present study, we addressed this topic in experiments which assess metacognitive judgments for audiovisual McGurk speech stimuli. In our first experiment, observers viewed nine different videos from three mutually exclusive categories: stimuli that are rarely integrated, stimuli that often produce a McGurk illusion, and fully congruent audiovisual stimuli. On each trial, they rated confidence from 0 (not at all confident) to 100 (extremely confident) in their perceptual judgment. Our hypothesis was that observers would display the lowest confidence in illusory McGurk trials, the highest confidence in

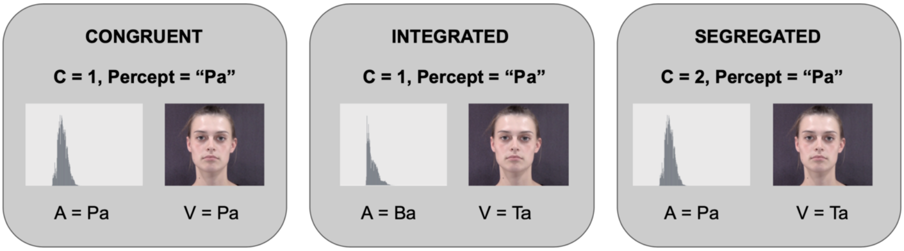

Figure 1. Different audiovisual speech–syllable combinations can yield identical reported percepts. In the standard McGurk paradigm (McGurk and MacDonald, 1976), an observer views a short audiovisual video which contains an auditory syllable (“A =” in each panel) and a visible person (or mouth) pronouncing a single syllable (“V =” in each panel). This paradigm can result in three scenarios: *congruent* speech, if the auditory and visual syllables are identical (left panel); *integrated* speech, if the auditory and visual syllables are inferred to originate from a single cause ($C = 1$) and produce perception of an intermediate, third syllable (middle panel); and *segregated* speech, where syllables differ enough that they are perceived as coming from distinct causes ($C = 2$) and the participant most often reports perceiving the auditory syllable (right panel). Interestingly, these three scenarios can potentially all give rise to the same perceived syllable (“Pa”), as in the example shown here.

fully congruent audiovisual speech, and moderate confidence in syllable combinations that are rarely integrated. In Experiment 2, we implemented a new task with 12 different audiovisual videos; our main analysis involved selecting pairs of trials where the integrated McGurk percept matched a particular type of congruent audiovisual trial (e.g., auditory Ba/visual Ga causes an observer to perceive "Da," and visual Da/auditory Da causes an observer to perceive "Da"). We hypothesized that observers would be more confident in fully congruent trials compared to McGurk trials, even when these trials resulted in the identical reports. In our third experiment, we evaluated whether confidence judgments would distinguish between scenarios where the reported percept was the same, but the underlying causal structure of the stimuli was different. Our hypothesis was that, on average, subjects would express higher confidence in stimuli that were rarely integrated, as these percepts would likely be less ambiguous than McGurk percepts. Thus, across our three experiments, we ask whether metacognition can differentiate between congruent, integrated, and segregated speech with unmatched reports (Experiment 1), whether metacognition can distinguish between congruent and integrated scenarios where the reported percept is the same and the underlying causal structure is the same ($C = 1$) (Experiment 2), and whether metacognition can distinguish between scenarios where the reported percept is the same, but the underlying causal structure is different (i.e., $C = 1$ and $\underline{C} = 2$) (Experiment 3).

Results from the first experiment revealed that confidence was highest on congruent speech, intermediate for rarely integrated incongruent speech, and lowest for McGurk speech, but a linear mixed-effects model (with speech condition as a fixed factor) revealed that the effect of speech type was not significant. In our second experiment, we showed that even when the perceived syllable was the same, confidence was higher for audiovisual congruent speech compared to McGurk speech. This indicates that even when multisensory percepts are identical (Deroy *et al.*, 2016; Freeman and Simoncelli, 2011), metacognition can distinguish congruent and integrated speech. In our third experiment, results showed that confidence judgments were quite similar between McGurk stimuli and rarely integrated stimuli with matched percepts. Together, these results provide evidence regarding both the capacities and limitations of metacognition to distinguish different types of multisensory information.

2. Experiment 1

2.1. Method

In Experiment 1, we explored whether confidence judgments for audiovisual speech differ across conditions with congruent audiovisual syllables, McGurk syllables, and rarely integrated audiovisual syllable combinations.

2.2. Participants

Thirty-five participants enrolled in this online experiment, which was coded in jsPsych 6.3.1 (de Leeuw, 2015), and launched using custom code through Google Drive's Application Programming Interface (API). Participants were recruited through Prolific.co, and were awarded $4.30 upon completion of the task. Three participants completed less than half of the task and were excluded from further analysis; additionally, two participants refreshed their browser window while completing the experiment and completed most of the task twice, so they were also excluded. Thus, 30 participants were included in our final analysis (18 men, 12 women; mean age = 32.8 years).

2.3. Stimuli

For our McGurk stimuli, we selected video and audio clips from the Oldenburg Audio Visual Speech Stimuli (OLAVS) set (Stropahl *et al.*, 2017) and created nine audiovisual combinations using Adobe Premiere Pro (Version 22) (see Table 1). Our reason for using these specific stimuli was to have a balanced stimulus design, with three fully-congruent audiovisual stimulus pairs (auditory Ba/visual Ba; auditory Ma /visual Ma; auditory Pa/visual Pa), three pairs known to produce illusory McGurk syllables (auditory Ba/visual Ga = perceive Da or Ma; auditory Pa/visual Na = perceive Ka or Ta, and auditory Ma/visual Ta = perceive Na or La), and three pairs of incongruent stimuli that are rarely integrated (auditory Da/visual Ma; auditory Na/visual Da; auditory Ta/visual Ga). All auditory and visual stimuli were from speaker TK01 in the Stropahl *et al.* (2017) dataset. There were four potential answer options on each trial. For McGurk trials, the possible answer choices corresponded to the visual syllable, the auditory syllable, and two syllables which reflect "fused" percepts (from Table 1 in Stropahl *et al.*, 2017). For Rarely Integrated trials, the possible answer choices corresponded to the visual syllable, the auditory

Table 1.
Audiovisual syllable combinations and answer options for Experiment 1

Stimulus type	Auditory syllable	Visual syllable	Answer options
Congruent	Ba	Ba	Ba, Pa, Ga, Ma
	Ma	Ma	Ma, Pa, Ga, Ba
	Pa	Pa	Pa, Ta, Ga, Ba
McGurk	Ba	Ga	Ba, Ga, Da, Ma
	Pa	Na	Pa, Na, Ka, Ta
	Ma	Ta	Ma, Ta, Na, La
Rarely Integrated	Da	Ma	Da, Ma, Ta, Ga
	Na	Da	Na, Da, Ka, Ma
	Ta	Ga	Ta, Ga, Ka, Da

syllable, and two “foil” answers that did not correspond to any presented or integrated percept. For congruent trials, we selected the correct syllable, as well as three potential foils.

2.4. *Procedure*

Participants began our online task by enrolling through the website Prolific.co. Following a welcome screen, participants read through our online consent form (IRB #201902462, University of Florida) and provided consent by checking a box next to the statement, “I agree to participate in this study.” Next, participants reported their sex, age, and viewed a photograph which demonstrated roughly how far they should be from the screen while participating in the experiment. Participants were then presented with a sample McGurk video, and were asked to adjust their speaker volume to a comfortable level. They could press a “repeat” button as many times as necessary to adjust the volume; a “continue” button moved the experiment forward.

Following this, participants completed nine practice trials. On each practice trial, a video of the speaker was presented for 2000 ms, and then the answer options were shown. Participants provided a categorical answer about the perceived syllable by pushing a button on the screen with their mouse cursor. Next, participants provided an answer about their confidence by moving a slider on the screen on a continuous scale from 0 (no confidence) to 100 (fully confident).

After the practice session, participants began the real experiment which consisted of three blocks of 45 trials. As in the practice, on each trial participants were shown an audiovisual video of a speaker, which could contain congruent, McGurk, or rarely integrated audiovisual syllables. Following each video, answer options were shown, which were customized for each specific video type (see Table 1). The answer option order was randomized on each trial. After they reported the perceived syllable, participants rated their confidence in each judgment on the scale from 0 to 100. The slider started from a random position on each trial, and moved in increments of 1 on the scale. Participants were allowed to take breaks between blocks. Upon finishing the task, participants were given a “completion code” to receive payment for their time and effort.

2.5. *Results*

We investigated whether metacognitive judgments could distinguish between McGurk speech and other forms of audiovisual speech. Use of a repeated-measures ANOVA in this design would produce biased standard errors and thus an increased type-I error rate, due to multiple ratings per subject and multiple ratings per item (Brauer and Curtin, 2018). Thus, to answer this question, we conducted a linear mixed-effects model with condition as a fixed factor with three levels (Congruent, McGurk, Rarely Integrated). Previous work advises that whenever a given subject provides multiple data points, a by-subject random intercept should be included, and when all subjects evaluate the same set of items, a by-item random intercept should be specified as well (Brauer and Curtin, 2018); thus, we also included random intercepts for subject and stimuli, as well as a by-subject random slope for our "condition" predictor, which varied within subjects.

We conducted our linear mixed-effects model in JAMOVI (Version 2.3.18.0; https://www.jamovi.org) and created plots using an identical model in Rstudio (2022.07.2; https://posit.co/download/rstudio-desktop). There was no effect of condition (see Table 2); specifically, the effects for McGurk–Congruent ($b = -8.21$, $t_{6.65} = -1.32$, $p = 0.23$) and Rarely Integrated–Congruent were not statistically significant ($b = -5.53$, $t_{7.09} = -0.87$, $p = 0.41$).

As shown in Fig. 2A, while the estimated marginal means revealed that the trend posited by our hypothesis was present, with Congruent having the highest confidence (Mean = 89.6, SE = 4.51), Rarely Integrated with second-highest confidence (Mean = 84.0, SE = 4.72), and McGurk stimuli with the lowest confidence (Mean = 81.4, SE = 4.85), heterogeneity existed in the random effects for each stimulus (Fig. 2B). For instance, even though congruent

Table 2.
Fixed-effect parameter estimates and random components for the linear mixed-effects model in Experiment 1

Fixed effects parament estimates

				95% confidence interval				
Names	Effect	Estimate	SE	Lower	Upper	df	t	p
(Intercept)	(Intercept)	84.99	2.96	79.2	90.79	11.80	28.714	<.001
Condition 1	McGurk–Congru	−8.21	6.22	−20.4	3.99	6.65	−1.319	0.231
Condition 2	Rarely–Congru	−5.53	6.32	−17.9	6.87	7.09	−0.874	0.411

Random components

Groups	Name	SD	Variance	ICC
Subject	(Intercept)	8.89	79.0	0.280
	Condition 1	7.67	58.8	
	Condition 2	9.85	97.0	
Stimulus	(Intercept)	7.39	54.7	0.212
Residual		14.27	203.6	

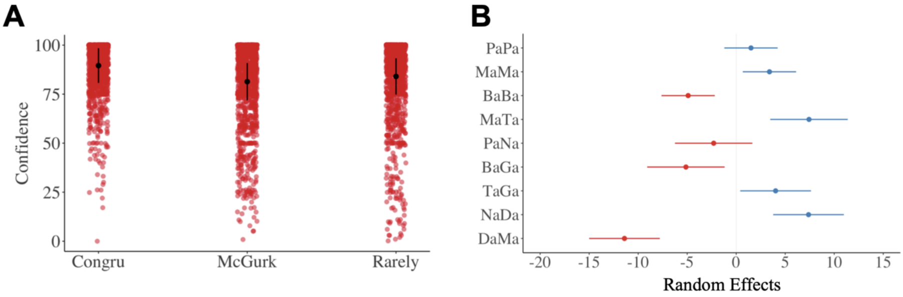

Figure 2. Predicted confidence ratings from our linear mixed-effects model and stimulus-specific random effects for judgments about audiovisual speech. (A) Predicted confidence ratings for stimuli from congruent, McGurk and Rarely Integrated conditions. (B) Random effects for each stimulus type. For each stimulus label, the auditory syllable is listed first, and the visual syllabus is listed second. In descending order, the first three labels are the congruent conditions; the next three labels are the McGurk conditions; the final three labels are the Rarely Integrated conditions.

trials had the highest overall confidence, the random effect for stimulus Auditory Ba–Visual Ba revealed a lower value than the random effect for stimulus Auditory Ma–Visual Ta, which was a McGurk stimulus. Variance for the intercepts across participants ($\sigma = 79.0$) and across stimuli ($\sigma = 54.7$) indicated the impact was notable [Intraclass Correlation Coefficient (ICC) = 0.28 and 0.21, respectively], but overall, speech type conditions had a minimal influence on confidence levels.

As expected, the number of syllable reports that selected the auditory stimulus exhibited high rates for the Congruent (mean = 92.7%, SD = 26.1%) and Rarely Integrated (mean = 93.1%, SD = 25.3%) conditions, but a low rate for selecting the auditory response in the McGurk condition (mean = 27.9%, SD = 44.9%). This low rate in the McGurk condition was primarily driven by a strong tendency to select one of the two perceptual fusion responses; on McGurk trials, fused responses made up 71.5% of answers, with responses to the auditory syllable much less frequent (27.9%) and responses reporting the visual component extremely rare (0.5%). As anticipated based on previous research (Stropahl *et al.*, 2017), the fusion rates for the three different McGurk conditions were quite comparable, with auditory Ba/visual Ga, auditory Pa/visual Na, and auditory Ma/visual Ta all yielding comparable rates of fused responses (means = 72.0%, 67.2%, and 75.2%, respectively).

These average binding rates across subjects provide one perspective on the data, but it is important to note that on a subject-by-subject basis, many subjects exhibited all-or-none binding with McGurk stimuli (Basu Mallick *et al.*, 2015), and some experienced binding in an intermediate range. For example, for auditory Ba/visual Ga McGurk stimuli, 12 out of 30 subjects always reported one of the two fused percepts, three never reported a fused percept, and 15 reported fused percepts for some proportion of the stimuli. For auditory Pa/visual Na McGurk stimuli, nine always reported a fused percept, five never reported a fused percept, and 16 reported fused percepts for some of the stimuli. Lastly, for auditory Ma/visual Ta stimuli, 16 always reported a fused percept, four never reported a fused percept, and 10 reported a fused percept at least some of the time.

3. Experiment 2

3.1. Method

After demonstrating in Experiment 1 that confidence was similar across congruent, McGurk, and rarely integrated speech, in Experiment 2, we aimed to determine whether confidence differed between congruent and illusory speech where the reported percept was identical (Deroy *et al.*, 2016). Previous research has identified pairs of stimuli which produce the same perceptual reports (Freeman and Simoncelli, 2011), but it remains an open question whether metacognitive systems distinguish between different types of multisensory stimuli yielding the same reported percept (Deroy *et al.*, 2016; Shams and Beierholm, 2021).

3.2. Participants

Forty-two participants enrolled in our second online experiment, which was coded using jsPsych 6.3.1 (de Leeuw, 2015) and Google's web API, and administered through Prolific.co. Participants were awarded $4.30 upon completion of the task. Seven participants completed less than half of the task and were excluded from further analysis; one participant refreshed the browser window while completing the experiment and completed most of the task twice, and was also excluded; two other subjects had data packet errors causing mismatches between trials and responses, and were also excluded. Thus, 32 participants were included in our final analysis (12 men, 19 women, one unreported; mean age = 29.9 years).

3.3. Stimuli

For our McGurk stimuli in Experiment 2, we selected video and audio clips from the Oldenburg Audio Visual Speech Stimuli (OLAVS) set (Stropahl *et al.*, 2017) and created twelve audiovisual combinations from Speaker TK01 using Adobe Premiere Pro (Version 22). Our reason for using stimuli from three conditions was to have a balanced stimulus design between fully congruent syllables, McGurk syllable combinations, and incongruent syllable combinations that are rarely integrated, so that participants would not always be integrating stimuli on every trial. Our fully congruent audiovisual stimulus pairs included auditory Na/visual Na, auditory Pa/visual Pa, auditory

Table 3.
Stimulus types, audiovisual syllable combinations, answer options, and the targeted matched response (between Congruent and McGurk conditions) for Experiment 2

Stimulus type	Auditory syllable	Visual syllable	Answer options	matched response
Congruent	Na	Na	Na, Ma, La, Ka	Na
	Pa	Pa	Pa, Ta, Ga, Ba	Pa
	Da	Da	Da, Ga, Ka, Ta	Da
	Ta	Ta	Ta, Ka, Ga, Da	Ta
McGurk	Ma	Ta	Ma, Ta, Na, La	Na
	Ba	Ta	Ba, Ta, Pa, Da	Pa
	Ba	Ka	Ba, Ka, Ga, Da	Da
	Pa	Da	Pa, Da, Ka, Ta	Ta
Rarely Integrated	Na	Da	Na, Da, Ka, Ma	N/A
	Pa	Ta	Pa, Ta, Ka, Da	N/A
	Ga	Ta	Ga, Ta, Ka, Na	N/A
	Ta	Ma	Ta, Ma, Pa, Da	N/A

Da/visual Da, and auditory Ta/visual Ta. Our McGurk stimulus pairs included auditory Ma/visual Ta, auditory Ba/visual Ta, auditory Ba/visual Ka, and auditory Pa/visual Da. Our incongruent pairs that were rarely integrated included auditory Na/visual Da, auditory Pa/visual Ta, auditory Ga/visual Ta, and auditory Ta/visual Ma (see Table 3). Our motivation for selecting these specific stimuli was to create conditions where the perceived syllable could be matched across the congruent and McGurk conditions. For example, the perceived syllable in each of the congruent conditions listed above is, respectively: Na, Pa, Da, Ta. After piloting the McGurk conditions, we found that the four McGurk conditions described above can also produce (in many individuals) the perceived syllables Na, Pa, Da, and Ta.

3.4. Procedure

Following acceptance of our online consent form, participants completed a “virtual chinrest” test to measure viewing distance, a sound check screen to allow them to adjust volume to a comfortable level, and introductory screens with task instructions. Subjects also reported their age and their biological sex on the introductory screens. Participants then completed three practice trials to acquaint themselves with the primary task. On each practice trial, a given video of a McGurk speaker was presented for approximately 2000 ms. They provided a categorical answer about the perceived syllable by pushing a button on the screen participants provided an answer about their confidence by moving a slider on a continuous scale from 0 (no confidence) to 100 (fully confident).

After the practice session, participants began the real experiment which consisted of 3 blocks of 60 trials. As in the first experiment, participants performed two tasks on each trial: (1) they reported the syllable they perceived after hearing/seeing the audiovisual video; and (2) they rated their confidence in each judgment on the scale from 0 to 100. Participants were allowed to take breaks between blocks. Upon finishing the task, participants were given a “completion code” to receive payment for their time and effort.

3.5. Results

To determine if confidence judgments could distinguish between congruent videos and McGurk videos when the reported syllable was the same, we first filtered our dataset to select responses where the condition was either Congruent or McGurk, and the perceived syllable was matched between those two conditions. Next, we fit a linear mixed-effects model with fixed effects for response, condition, and the response*condition interaction, as well as random intercepts for subjects and stimuli. Results of our fixed-effects model are shown below in Table 4.

However, we note that our primary comparison of interest for this task was to evaluate if confidence differed between congruent and illusory speech where the reported percept was identical. The *post-hoc* test for the difference between Congruent and McGurk stimuli with matched percepts was significant ($t = 2.13$, $p = 0.03$); the estimated marginal means showed that overall, with matched syllable reports, the congruent stimuli exhibited higher overall confidence (mean = 91.8, SE = 3.28) than the McGurk stimuli (mean = 82.9, SE = 3.54), and this trend appeared to be evident across all four response types: Da, Na, Pa, and Ta (Fig. 3). To further explore whether the differences between congruent and McGurk responses to any specific response type were significant, we investigated the simple effects of condition; interestingly, none of the four tests were significant (Da: $t = -1.51$, $p = 0.13$; Na: $t = -0.56$, $p = 0.58$; Pa: $t = -0.84$, $p = 0.40$; Ta: $t = -1.41$, $p = 0.16$). This was driven in large part to include the random intercept for stimulus in our model; when this intercept was removed, all *post-hoc* tests were highly significant, but since each subject evaluated the same set of items (Brauer and Curtin, 2018), we concluded that it was best to include this term.

Overall, the rates of how often participants perceived “fused” stimuli in the McGurk conditions was 71.4%, with fusion rates ranging between 61% and 79% for any specific condition. This fusion rate is in line with previous reports using these videos (Stropahl *et al.*, 2017). The next most common responses in the McGurk trials were to report the auditory component of what was shown (26.0%), and the least common response was to report the visual component of the video (2.5%).

Table 4.
Fixed-effect parameter estimates and random components for the linear mixed-effects model in Experiment 2

Fixed effects parament estimates

				95% confidence interval				
Names	Effect	Estimate	SE	Lower	Upper	df	t	p
(Intercept)	(Intercept)	87.358	2.70	82.08	95.640	183	32.4146	<.001
Response 1	Na–Da	7.739	5.63	−3.29	18.764	2426	1.3758	0.169
Response 2	Pa–Da	5.023	6.21	−7.14	17.186	2427	0.8095	0.418
Response 3	Ta–Da	5.301	5.62	−5.72	16.324	2426	0.9426	0.346
Condition 1	McGurk–Congru	−8.938	4.19	−17.15	−0.731	2426	−2.1345	0.033
Response 1 * Condition 1	Na–Da*McGurk–Congru	7.573	11.25	−14.48	29.622	2426	0.6732	0.501
Response 2 * Condition 1	Pa–Da*McGurk–Congru	3.955	12.41	−20.36	28.273	2426	0.3188	0.750
Response 3 * Condition 1	Ta–Da*McGurk–Congru	0.842	11.25	−21.20	22.887	2426	0.0749	.0940

Random components

Groups	Name	SD	Variance	ICC
Subject	(Intercept)	9.60	92.1	0.359
Stimulus	(Intercept)	5.58	31.1	0.159
Residual		12.82	164.3	

On a subject-by-subject basis, many subjects exhibited all-or-none binding with McGurk stimuli, and some experienced binding in an intermediate range. For example, for auditory Ma/visual Ta McGurk stimuli, 12 out of 32 subjects always reported one of the two fused percepts, two never reported a fused percept, and 18 reported fused percepts for some proportion of the stimuli. For auditory Ba/visual Ta McGurk stimuli, 12 always reported a fused percept, one never reported a fused percept, and 19 reported fused percepts for some of the stimuli. For auditory Ba/visual Ka McGurk stimuli, 12 always reported a fused percept, one never reported a fused percept, and 19 reported fused percepts for some of the stimuli. Lastly, for auditory Pa/visual Da stimuli, nine always

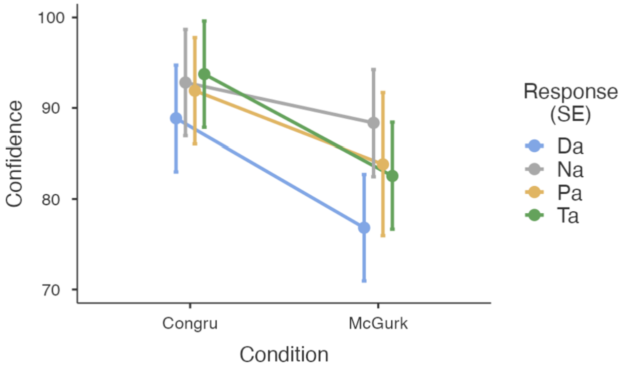

Figure 3. Confidence in audiovisual speech perception for matched syllables. The estimated marginal means are plotted for specific response types (Da, Na, Pa, Ta) for the congruent and McGurk conditions. Error bars represent standard error.

reported a fused percept, three never reported a fused percept, and 20 reported a fused percept at least some of the time.

4. Experiment 3

4.1. Method

In Experiment 3, we aimed to determine whether confidence differs between McGurk speech and Rarely Integrated speech when the reported percept was identical (Deroy *et al.*, 2016). While Experiment 2 provided evidence that confidence distinguished between Congruent and illusory McGurk speech with matched percepts, both of these conditions are characterized by the same causal scenario (Fig. 1). Thus, it remains an open question whether metacognitive systems distinguish between different types of multisensory stimuli when the causal structure differs across stimuli with matched reports (Deroy *et al.*, 2016; Shams and Beierholm, 2021).

4.2. Participants

Thirty participants enrolled in this online experiment, which was coded in jsPsych 6.3.1 (de Leeuw, 2015), and launched through the website Cognition.run. Participants were recruited through Prolific.co and were awarded $4.30 upon completion of the task. All participants successfully completed the task and no data were excluded; thus, 30 participants were included in our final analysis (24 women, six men; mean age = 34.3 years).

4.3. Stimuli

For our McGurk stimuli in Experiment 3, we again utilized video and audio clips from the Oldenburg Audio Visual Speech Stimuli (OLAVS) (Stropahl

Table 5.
Stimulus types, audiovisual syllable combinations, answer options, and the possible matched syllable (between the McGurk and Rarely Integrated conditions) for Experiment 3

Stimulus type	Auditory syllable	Visual syllable	Answer options	Matched response
McGurk	Ma	Ta	Ma, Ta, Na, La	Na
	Ba	Ka	Ba, Ka, Ga, Da	Da
	Pa	Da	Pa, Da, Ka, Ta	Ta
Rarely Integrated	Na	Da	Na, Da, Ka, Ma	Na
	Da	Ma	Da, Ma, Ba, Na	Da
	Ta	Ga	Ta, Ga, Ka, Da	Ta

et al., 2017) and created six audiovisual combinations from Speaker TK01 using Adobe Premiere Pro (Version 22). We had two conditions: McGurk stimuli and “Rarely Integrated” stimuli, with each condition consisting of three unique audiovisual combinations. The McGurk stimuli pairs included auditory Ma/visual Ta, auditory Ba/visual Ka, and auditory Pa/visual Da. The “Rarely Integrated” incongruent stimulus pairs included auditory Na/visual Da, auditory Da/visual Ma, and auditory Ta/visual Ga. Our motivation for selecting these specific stimuli was to create conditions with matched reported percepts between the McGurk and Rarely Integrated conditions, to evaluate how metacognition tracks cue recovery. Thus, the anticipated perceived syllables for each of the three McGurk and Rarely Integrated conditions were “Na,” “Da,” and “Ta,” respectively (Table 5).

For McGurk trials, the possible answer choices that were shown to subjects corresponded to the visual syllable, the auditory syllable, and two syllables which reflect “fused” percepts (from Stropahl *et al.*, 2017). For Rarely Integrated trials, the possible answer choices corresponded to the visual syllable, the auditory syllable, and two “foil” answers that did not correspond to any presented or integrated percept.

4.4. Procedure

As in Experiment 1, participants began our online task by enrolling through the website Prolific.co. Following a welcome screen, participants read through our online consent form (IRB #201902462, University of Florida) and provided consent by checking a box next to the statement, “I agree to participate in this study.” Next, participants viewed an example photo which demonstrated how they should sit roughly one arm’s length away from the screen. Following this, each participant’s web camera was activated and they needed to ensure that their face was within view of the camera. After reporting their sex and age, participants were shown a sample 2-s McGurk stimulus video, and instructed that they could adjust the volume to a comfortable level for the experiment.

Participants could repeat the video as many times as necessary as they adjusted the volume.

Next, participants completed six practice trials, viewing each of the six stimuli one time. On each trial, participants reported which syllable they perceived (selected from one of four possible options), and then rated their confidence in this report by moving a slider on a continuous scale from 0 (no confidence) to 100 (fully confident). Once the practice trials finished, participants began the main experiment, which consisted of two blocks of 60 trials (120 trials totals). Each of the three McGurk stimuli were presented 20 times each in the experiment; each of the three rarely integrated stimuli were presented 20 times each in the experiment. As in the practice session, participants reported the syllable on each trial, and also rated their confidence on the 0 to 100 scale. Upon completing the task, participants were given a completion code and were re-routed to Prolific.co to receive payment.

4.5. Results

To determine if confidence judgments could distinguish between McGurk videos and Rarely Integrated videos when the reported syllable was the same, we first filtered our dataset to select responses where the condition was either McGurk or Rarely Integrated, and the perceived syllable was the one that we aimed to match between those two conditions (see Table 5). Next, we fit a linear mixed-effects model with fixed effects for response, condition, and the response*condition interaction, as well as random intercepts for subjects and stimuli. None of the fixed effects in this model were significant (see Table 6).

The estimated marginal means showed that, with matched syllable reports, the McGurk stimuli exhibited very similar overall confidence (mean = 87.6, SE = 4.34) to the Rarely Integrated stimuli (mean = 88.6, SE = 4.34), though we do note some stimulus-specific heterogeneity (see Fig. 4). The estimated marginal mean was slightly higher for rarely integrated “Ta” responses (mean = 92.4; SE = 6.86) compared to McGurk “Ta” responses (mean = 90.0; SE = 6.87) and slightly higher for rarely integrated “Na” responses (mean = 91.4; SE = 6.86) compared to McGurk “Na” responses (mean = 88.0; SE = 6.86), but simple-effects tests confirmed that neither of these differences were significant (Ta: $t = 0.27$, $p = 1.0$; Na: $t = 0.37$, $p = 1.0$). The estimated marginal mean for Da responses was slightly higher for McGurk (mean = 84.9; SE = 6.87) compared to Rarely Integrated stimuli (mean = 82.0; SE = 6.86), but this difference was also not statistically significant ($t = -0.31$, $p = 1.0$).

Because there were two answer choices on each McGurk trial that could be considered “fused,” (see Table 1; Stropahl *et al.*, 2017) this paradigm carried the risk of not producing matched syllable content. While three of our four conditions were quite robust in producing high numbers of subjects with

Table 6.
Fixed-effect parameter estimates and random components for the linear mixed-effects model in Experiment 3

Fixed effects parament estimates

				95% confidence interval				
Names	Effect	Estimate	SE	Lower	Upper	df	t	p
(Intercept)	(Intercept)	88.100	3.43	81.37	94.8	1.19e−7	25.660	1.000
Response 1	Na–Da	6.283	6.51	−6.46	19.0	4.19e−8	0.967	1.000
Response 2	Ta–Da	7.774	6.51	−4.98	20.5	4.19e−8	1.194	1.000
Condition 1	Rarely–McGurk	0.996	5.32	−9.42	11.4	4.19e−8	0.187	1.000
Response 1*Condition 1	Na–Da*Rarely–McGurk	6.300	13.02	−19.21	31.8	4.19e−8	0.484	1.000
Response 2*Condition 1	Ta–Da*Rarely–McGurk	5.344	13.02	−20.17	30.9	4.19e−8	0.410	1.000

Random components

Groups	Name	SD	Variance	ICC
Subject	(Intercept)	11.91	141.8	0.557
Stimulus	(Intercept)	6.49	42.1	0.272
Residual		10.61	112.6	

matched syllable content, we did notice that one condition was deficient in this sample: for auditory Ba/visual Ta, perceiving the McGurk syllable “Pa” was a relative rarity, with only four out of 32 subjects producing matched syllable content. This contrasted the robust matching in the other three conditions, with each condition producing matched syllable content in at least 29 out of the 32 subjects.

Overall, the rates of how often participants perceived “fused” stimuli in the McGurk conditions was 64.9%, with fusion rates ranging between 61% and 69% for any specific condition (Ba–Ka mean fusion = 68.5%, SD = 46.5%; Ma–Ta mean fusion = 64.7% SD = 47.8%, Pa–Da mean fusion = 61.5%, SD = 48.7%). This fusion rate is in line with previous reports using these videos (Stropahl *et al.*, 2017). The next most common responses in the McGurk trials were to report the auditory component of what was shown

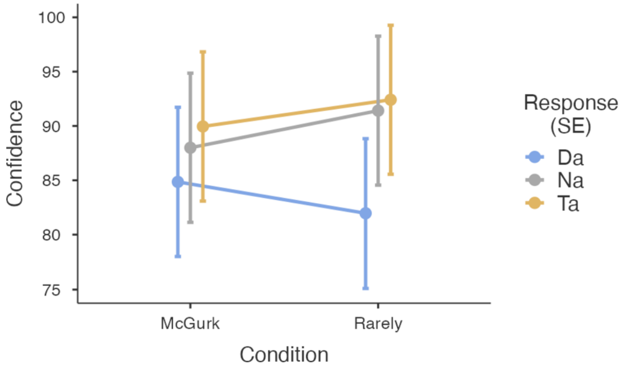

Figure 4. Confidence in audiovisual speech perception for McGurk and Rarely Integrated with matched perceived syllables. The estimated marginal means are plotted for specific response types (Da, Na, Ta) for the McGurk and Rarely Integrated conditions. Error bars represent standard error.

(34.0%), and the least common response was to report the visual component of the video (1.1%).

On a subject-by-subject basis, many subjects exhibited all-or-none binding with McGurk stimuli (Basu Mallick *et al.*, 2015), and some experienced binding in an intermediate range. For example, for auditory Ma/visual Ta McGurk stimuli, 11 out of 30 subjects always reported one of the two fused percepts, four never reported a fused percept, and 15 reported fused percepts for some proportion of the stimuli. For auditory Ba/visual Ka McGurk stimuli, 11 always reported a fused percept, four never reported a fused percept, and 15 reported fused percepts for some of the stimuli. Lastly, for auditory Pa/visual Da stimuli, eight always reported a fused percept, six never reported a fused percept, and 16 reported a fused percept at least some of the time.

5. Discussion

In this investigation, we aimed to answer two questions: can metacognition distinguish between congruent, integrated, and segregated audiovisual speech? And can it do so when the reported percepts are the same? Using McGurk speech stimuli (Stropahl *et al.*, 2017), in our first experiment, results showed that, despite slight differences, confidence judgments were quite similar overall between Congruent speech, McGurk speech, and Rarely Integrated speech. In our second experiment, when the perceptual content was matched between Congruent and McGurk trials, confidence was higher for each congruent condition compared to the matched McGurk percept. Our third experiment showed that confidence in perceived syllables for Rarely Integrated trials and McGurk trials with matched percepts were quite similar. These results indicate that it might be challenging for confidence judgments to distinguish between different causal structures when the percept is matched.

As previous research has noted (Deroy *et al.*, 2016), little is currently known about the relationship between metacognition and multisensory integration (Garzorz and Deroy, 2020). One previous investigation provided preliminary evidence that confidence differs between different forms of audiovisual speech: in a study of speech perception in schizophrenia, age-matched controls exhibited lower confidence (on average) with McGurk speech, compared to fully congruent or incongruent audiovisual speech (White *et al.*, 2014). Results from our first experiment show that rates of confidence between Congruent, McGurk, and Rarely Integrated speech are quite similar. Presently, questions about whether or not metacognition can distinguish between stimuli that yield matched reports with different underlying causal structures, or access the underlying unisensory cues in different contexts, have remained unanswered (Deroy *et al.*, 2016; Shams and Beierholm, 2021). Here, we provide preliminary evidence that even when reported percepts are identical, confidence judgments can distinguish between two specific speech types where the inferred causal structure is the same (Congruent and McGurk stimuli), but confidence judgments may struggle to tease apart McGurk and Rarely Integrated stimuli, where the inferred causal structure is different.

Our findings can inform the debate about whether metacognition is involved in perceptual reality monitoring (Gershman, 2019; Lau, 2019), as higher-order systems may assist in discriminating between potential sources of sensory experiences. This can involve discriminations such as determining whether or not perceptual experiences are imagined or real (Dijkstra *et al.*, 2017, 2018, 2019, 2021), or in this case, determining the final percept for Congruent, McGurk, and Rarely Integrated speech. Interestingly, it appears these metacognitive systems can make fine-grained discriminations between matched percepts involving multisensory illusions and fully congruent stimuli (van Erp *et al.*, 2013). Thus, our findings here provide evidence of the abilities (and inabilities) of metacognition in being able to distinguish different sources of sensory experience, across a range of internally and externally generated sources.

One question that is not answered by the current investigation is whether our findings here would generalize to paradigms where the sensory signals obtained by observers are less salient. Specifically, multisensory speech stimuli include signals that are easy to perceive; the observer's lips and voice, at least in this study, are not particularly "noisy" sources of information (Bishop and Miller, 2009). Whether our findings would generalize to paradigms that include more ambiguous stimuli remains to be seen. For example, stimuli that are frequently used to study spatial and temporal interactions in multisensory paradigms are often much more ambiguous (Alais and Burr, 2004; Bertelson and Radeau, 1976; Chen and Vroomen, 2013; Ernst and Banks, 2002; Parise *et al.*, 2012; Shams *et al.*, 2000; Welch and Warren, 1980). Thus, future studies should aim to see whether results from these experiments replicate in other multisensory paradigms, including speech paradigms which use different speakers and syllable combinations than the ones employed in this study.

One way to improve upon the present design is to sample a much wider range of audiovisual syllable combinations when comparing across Congruent, McGurk, and Rarely Integrated speech. In the literature on mixed-effects models (e.g., Judd *et al.*, 2012), it is recommended to sample a wide and representative range of stimuli for whatever conditions will be compared against one another. Therefore, moving forward, as studies of multisensory integration increasingly incorporate this powerful statistical technique, we recommend expanding the number of stimuli incorporated into specific conditions to increase the power to detect smaller effects that may be present. We also recommend expanding the number of participants used in studies of this topic. For instance, it remains possible that the current study was underpowered to detect an effect in Experiment 1. It also is possible that McGurk rates differ across cultures; future studies should aim to systematically explore differences in McGurk rates across countries and the degree to which English is a first, second, third (or higher) language.

Further, we think it is important to explore additional "Rarely Integrated" cases where there is no perceived fusion, but there are strong conflicts that are difficult to reconcile. For example, in the case of auditory Ba and visual Fa, there is no integration, but there may be clear representations of both the auditory perceived syllable, as well as the visual syllable (as Fa has such a distinct appearance during lip movements). This case stresses the importance of dual-response paradigms to determine under which cases participants can accurately represent both the auditory and the visual components of auditory speech, separate from any intermediate percept that may emerge. These types of stimuli may also reveal interesting cases of complete visual capture of auditory perception.

Moving forward, paradigms can expand beyond asking about the syllable and corresponding confidence judgment to make explicit judgments of causal inference. For instance, it would be interesting to distinguish between three different types of potential responses: asking about the reported percept on a given trial; asking if the number of causes is one or two on a given trial; and asking about the probability of a common cause on a specific trial. Distinguishing these three constructs is critical because a fusion response (or auditory-only response) can occur even if $p_{(C=1)}$ is low because enough weight is given to the $C = 1$ percept to drag the final representation into a specific region of syllable space. Similarly, a fusion response may not occur even if $p_{(C\,=\,1)}$ is much higher if the integrated percept is still within the “ba” region of space.

Overall, we think these results can inform recent work which characterizes perceptual awareness as a higher-order state in generative models of perceptual contents (Fleming, 2020). Specifically, while the Bayesian Causal Inference models that have dominated multisensory research for the last decade and a half have produced profound insights into both behavioral (Körding *et al.*, 2007; Samad *et al.*, 2015) and neural (Rideaux *et al.*, 2021; Rohe and Noppeney, 2015; Rohe *et al.*, 2019) correlates of multisensory integration, linking Bayesian models to sensory phenomenology and sensory awareness has proven extremely difficult (Denison *et al.*, 2020). Our results provide preliminary evidence that metacognition may help distinguish between different sources of stimuli, even when the reported percept is the same. Future research should aim to explore the degree to which metacognition helps us make sense of both the internal and external origins of sensory causes.

Acknowledgements

Victoria Cardenas coded Experiment 1 and assisted with data collection using Google Drive’s webAPI and subject payments *via* Prolific.co. Samantha Pedersen coded Experiment 2 and assisted with data collection using Google Drive’s webAPI and subject payments *via* Prolific.co. Victoria Cardenas and Camila Rubiera coded Experiment 3 and assisted with data collection using Google Drive’s webAPI and subject payments *via* Prolific.co. Faith Kimmet coded Experiment 3, assisted with data collection on Cognition.run and subject payments *via* Prolific.co. Grey Johnson assisted with the online implementation of Experiments 1, 2, and 3. Brian Odegaard coded the analyses for Experiments 1–3. Addison Sans proofread the manuscript, supervised online data collection, and provided feedback on all code. Brian Odegaard, Faith Kimmet, and Addison Sans wrote the manuscript jointly, which was proofread and endorsed by all authors. Matt Baldwin assisted with the implementation of linear mixed-effects models, and reviewed the justification and statistical interpretation of results. Upon manuscript acceptance, all code and analyses for this project will be released on the Open Science Framework.

References

Alais, D. and Burr, D. (2004). The ventriloquist effect results from near-optimal bimodal integration, *Curr. Biol.* **14**, 257–262. DOI:10.1016/j.cub.2004.01.029.

Basu Mallick, D., Magnotti, J. F. and Beauchamp, M. S. (2015). Variability and stability in the McGurk effect: contributions of participants, stimuli, time, and response type, *Psychon. Bull. Rev.* **22**, 1299–1307. DOI:10.3758/s13423-015-0817-4.

Bertelson, P. and Radeau, M. (1976). Ventriloquism, sensory interaction, and response bias: remarks on the paper by Choe, Welch, Gilford, and Juola, *Percept. Psychophys.* **19**, 531–535. DOI:10.3758/BF03211222.

Bishop, C. W. and Miller, L. M. (2009). A multisensory cortical network for understanding speech in noise, *J. Cogn. Neurosci.* **21**, 1790–1805. DOI:10.1162/jocn.2009.21118.

Brauer, M. and Curtin, J. J. (2018). Linear mixed-effects models and the analysis of nonindependent data: a unified framework to analyze categorical and continuous independent variables that vary within-subjects and/or within-items, *Psychol. Methods* **23**, 389–411. DOI:10.1037/met0000159.

Chen, L. and Vroomen, J. (2013). Intersensory binding across space and time: a tutorial review, *Atten. Percept. Psychophys.* **75**, 790–811. DOI:10.3758/s13414-013-0475-4.

de Leeuw, J. R. (2015). jsPsych: a JavaScript library for creating behavioral experiments in a web browser, *Behav. Res. Methods* **47**, 1–12. DOI:10.3758/s13428-014-0458-y.

Denison, R. N., Block, N. and Samaha, J. (2020). What do models of visual perception tell us about visual phenomenology? *PsyArchiv*. DOI:10.31234/osf.io/7p8jg.

Deroy, O., Spence, C. and Noppeney, U. (2016). Metacognition in multisensory perception, *Trends Cogn. Sci.* **20**, 736–747. DOI:10.1016/j.tics.2016.08.006.

Dijkstra, N., Bosch, S. E. and van Gerven, M. A. J. (2017). Vividness of visual imagery depends on the neural overlap with perception in visual areas, *J. Neurosci.* **37**, 1367–1373. DOI:10.1523/JNEUROSCI.3022-16.2016.

Dijkstra, N., Mostert, P., Lange, F. P. de, Bosch, S. and van Gerven, M. A. (2018). Differential temporal dynamics during visual imagery and perception, *Elife* **7**, e33904. DOI:10.7554/eLife.33904.

Dijkstra, N., Bosch, S. E. and van Gerven, M. A. J. (2019). Shared neural mechanisms of visual perception and imagery, *Trends Cogn. Sci.* **23**, 423–434. DOI:10.1016/j.tics.2019.02.004.

Dijkstra, N., Kok, P. and Fleming, S. M. (2021). Perceptual reality monitoring: neural mechanisms dissociating imagination from reality, *PsyArchiv*. DOI:10.31234/osf.io/zngeq.

Ernst, M. O. and Banks, M. S. (2002). Humans integrate visual and haptic information in a statistically optimal fashion, *Nature* **415**, 429–433. DOI:10.1038/415429a.

Flavell, J. H. (1979). Metacognition and cognitive monitoring: a new area of cognitive–developmental inquiry, *Am Psychol.* **34**, 906–911. DOI:10.1037/0003-066X.34.10.906.

Fleming, S. M. (2020). Awareness as inference in a higher-order state space, *Neurosci. Conscious.* **2020**, niz020. DOI:10.1093/nc/niz020.

Fleming, S. M. and Lau, H. C. (2014). How to measure metacognition, *Front. Hum. Neurosci.* **8**, 443. DOI:10.3389/fnhum.2014.00443.

Freeman, J. and Simoncelli, E. P. (2011). Metamers of the ventral stream, *Nat. Neurosci.* **14**, 1195–1201. DOI:10.1038/nn.2889.

Garzorz, I. and Deroy, O. (2020). Why there is a vestibular sense, or how metacognition individuates the senses, *Multisens. Res.* **34**, 261–280. DOI:10.1163/22134808-bja10026.

Gershman, S. J. (2019). The generative adversarial brain, *Front. Artif. Intell.* **2**, 18. DOI:10.3389/frai.2019.00018.

Judd, C. M., Westfall, J. and Kenny, D. A. (2012). Treating stimuli as a random factor in social psychology: a new and comprehensive solution to a pervasive but largely ignored problem, *J. Pers. Soc. Psychol.* **103**, 54–69. DOI:10.1037/a0028347.

Knill, D. C. and Richards, W. (1996). *Perception as Bayesian Inference*. Cambridge University Press, Cambridge, UK. DOI:10.1017/CBO9780511984037.

Körding, K. P., Beierholm, U., Ma, W. J., Quartz, S., Tenenbaum, J. B. and Shams, L. (2007). Causal inference in multisensory perception, *PloS ONE* **2**, e943. DOI:10.1371/journal.pone.0000943.

Lau, H. (2019). Consciousness, metacognition and perceptual reality monitoring, *PsyArchiv*. DOI:10.31234/osf.io/ckbyf.

Magnotti, J. F. and Beauchamp, M. S. (2017). A causal inference model explains perception of the McGurk effect and other incongruent audiovisual speech, *PLoS Comput. Biol.* **13**, e1005229. DOI:10.1371/journal.pcbi.1005229.

Magnotti, J. F., Ma, W. J. and Beauchamp, M. S. (2013). Causal inference of asynchronous audiovisual speech, *Front. Psychol.* **4**, 798. DOI:10.3389/fpsyg.2013.00798.

Magnotti, J. F., Smith, K. B., Salinas, M., Mays, J., Zhu, L. L. and Beauchamp, M. S. (2018). A causal inference explanation for enhancement of multisensory integration by co-articulation, *Sci. Rep.* **8**, 18032. DOI:10.1038/s41598-018-36772-8.

McGovern, D. P., Roudaia, E., Newell, F. N. and Roach, N. W. (2016). Perceptual learning shapes multisensory causal inference via two distinct mechanisms, *Sci. Rep.* **6**, 24673. DOI:10.1038/srep24673.

McGurk, H. and MacDonald, J. (1976). Hearing lips and seeing voices, *Nature* **264**, 746–748. DOI:10.1038/264746a0.

Odegaard, B., Wozny, D. R. and Shams, L. (2015). Biases in visual, auditory, and audiovisual perception of space, *PLoS Comput. Biol.* **11**, e1004649. DOI:10.1371/journal.pcbi.1004649.

Parise, C. V., Spence, C. and Ernst, M. O. (2012). When correlation implies causation in multisensory integration, *Curr. Biol.* **22**, 46–49. DOI:10.1016/j.cub.2011.11.039.

Pick, H. L., Warren, D. H. and Hay, J. C. (1969). Sensory conflict in judgments of spatial direction, *Percept. Psychophys.* **6**, 203–205. DOI:10.3758/BF03207017.

Rahnev, D., Balsdon, T., Charles, L., de Gardelle, V., Denison, R., Desender, K., Faivre, N., Filevich, E., Fleming, S. M., Jehee, J., Lau, H., Lee, A. L. F., Locke, S. M., Mamassian, P., Odegaard, B., Peters, M., Reyes, G., Rouault, M., Sackur, J., Samaha, J., Sergent, C., Sherman, M. T., Siedlecka, M., Soto, D., Vlassova, A. and Zylberberg, A. (2022). Consensus goals in the field of visual metacognition, *PsyArchiv*. DOI:10.31234/osf.io/z8v5x.

Rideaux, R., Storrs, K. R., Maiello, G. and Welchman, A. E. (2021). How multisensory neurons solve causal inference, *Proc. Natl Acad. Sci. U. S. A.* **118**. DOI:10.1073/pnas.2106235118.

Rohe, T. and Noppeney, U. (2015). Cortical hierarchies perform Bayesian causal inference in multisensory perception, *PLoS Biol.* **13**, e1002073. DOI:10.1371/journal.pbio.1002073.

Rohe, T., Ehlis, A.-C. and Noppeney, U. (2019). The neural dynamics of hierarchical Bayesian causal inference in multisensory perception, *Nat. Commun.* **10**, 1907. DOI:10.1038/s41467-019-09664-2.

Samad, M., Chung, A. J. and Shams, L. (2015). Perception of body ownership is driven by Bayesian sensory inference, *PloS ONE* **10**, e0117178. DOI:10.1371/journal.pone.0117178.

Shams, L. and Beierholm, U. R. (2010). Causal inference in perception, *Trends Cogn. Sci.* **14**, 425–432.

Shams, L. and Beierholm, U. (2021). Bayesian causal inference: a unifying neuroscience theory, *PsyArchiv*. DOI:10.31234/osf.io/xpz6n.

Shams, L., Kamitani, Y. and Shimojo, S. (2000). Illusions. What you see is what you hear, *Nature* **408**, 788. DOI:10.1038/35048669.

Stevenson, R. A., Zemtsov, R. K. and Wallace, M. T. (2012). Individual differences in the multisensory temporal binding window predict susceptibility to audiovisual illusions, *J. Exp.*

Psychol. Hum. Percept. Perform. J. Exp. Psychol. Hum. Percept. Perform. **38**, 1517–1529. DOI:10.1037/a0027339.

Stropahl, M., Schellhardt, S. and Debener, S. (2017). McGurk stimuli for the investigation of multisensory integration in cochlear implant users: the Oldenburg Audio Visual Speech Stimuli (OLAVS), *Psychon. Bull. Rev.* **24**, 863–872. DOI:10.3758/s13423-016-1148-9.

van Erp, J. B. F., Philippi, T. G. and Werkhoven, P. (2013). Observers can reliably identify illusory flashes in the illusory flash paradigm, *Exp. Brain Res.* **226**, 73–79. DOI:10.1007/s00221-013-3413-8.

Welch, R. B. and Warren, D. H. (1980). Immediate perceptual response to intersensory discrepancy, *Psychol. Bull.* **88**, 638–667.

White, T. P., Wigton, R. L., Joyce, D. W., Bobin, T., Ferragamo, C., Wasim, N., Lisk, S. and Shergill, S. S. (2014). Eluding the illusion? Schizophrenia, dopamine and the McGurk effect, *Front. Hum. Neurosci.* **8**, 565. DOI:10.3389/fnhum.2014.00565.