Attention, Perception, & Psychophysics (2021) 83:2583–2598
https://doi.org/10.3758/s13414-021-02265-6

Rethinking the McGurk effect as a perceptual illusion

Laura M. Getz[1,2] · Joseph C. Toscano[1]

Accepted: 25 January 2021 / Published online: 21 April 2021

Abstract

Visual speech cues play an important role in speech recognition, and the McGurk effect is a classic demonstration of this. In the original McGurk & Macdonald (*Nature 264*, 746–748 1976) experiment, 98% of participants reported an illusory "fusion" percept of /d/ when listening to the spoken syllable /b/ and watching the visual speech movements for /g/. However, more recent work shows that subject and task differences influence the proportion of fusion responses. In the current study, we varied task (forced-choice vs. open-ended), stimulus set (including /d/ exemplars vs. not), and data collection environment (lab vs. Mechanical Turk) to investigate the robustness of the McGurk effect. Across experiments, using the same stimuli to elicit the McGurk effect, we found fusion responses ranging from 10% to 60%, thus showing large variability in the likelihood of experiencing the McGurk effect across factors that are unrelated to the perceptual information provided by the stimuli. Rather than a robust perceptual illusion, we therefore argue that the McGurk effect exists only for some individuals under specific task situations.

Significance: This series of studies re-evaluates the classic McGurk effect, which shows the relevance of visual cues on speech perception. We highlight the importance of taking into account subject variables and task differences, and challenge future researchers to think carefully about the perceptual basis of the McGurk effect, how it is defined, and what it can tell us about audiovisual integration in speech.

Keywords McGurk effect · Audiovisual integration · Speech perception

Speech perception is a multimodal process (Rosenblum, 2008), such that visual information from the speaker's mouth movements provides synchronous and redundant information to the acoustic signals heard during normal face-to-face conversations. This integrated audiovisual (AV) percept has been shown to aid spoken language comprehension by adults in both quiet (McGettigan et al., 2012; Sánchez-García, Alsius, Enns, & Soto-Faraco, 2011)and for speech in background noise (Gilbert, Lansing, & Garnsey, 2012; Ross, Saint-Amour, Leavitt, Javitt, & Foxe, 2007; Sumby & Pollack, 1954; Summerfield, 1979).

A critical question concerns precisely how listeners integrate auditory and visual information into a unified percept. This ability emerges over the course of development (Sekiyama & Burnham, 2004; Knowland, Mercure, Karmiloff-Smith, Dick, & Thomas, 2014), with evidence from infants suggesting that they are sensitive to auditory and visual speech characteristics even before they begin producing words (Rosenblum, 2008). Further, infants can detect temporal (Lewkowicz, 2010; Pons & Lewkowicz, 2014; Dodd, 1979) and phonemic (Kuhl & Meltzoff, 1984; Patterson & Werker, 1999; Patterson & Werker, 2003; Aldridge, Braga, Walton, & Bower, 1999; Kitamura, Guellaï, & Kim, 2014; Guellaï, Streri, Chopin, Rider, & Kitamura, 2016)asynchrony in isolated syllables and continuous speech, providing evidence that tracking these co-occurrences develops rapidly. However, despite early sensitivity to AV synchrony, children do not fully integrate multimodal speech cues until later in development. Children weight visual cues less compared to adults in speech categorization tasks involving cues from both modalities (Massaro, Thompson, Barron, & Laren, 1986; Hirst, Stacey, Cragg, Stacey, & Allen, 2018), even in cases where the auditory signal is degraded (Maidment, Kang, Stewart, & Amitay, 2015; Wightman, Kistler, & Brungart, 2006; Barutchu et al.,

✉ Laura M. Getz
lgetz@sandiego.edu

[1] Department of Psychological and Brain Sciences, Villanova University, Villanova, PA USA

[2] Department of Psychological Sciences, University of San Diego, 5998 Alcala Park, San Diego, CA, 92110, USA

2010; Ross et al., 2011). These studies suggest that the accuracy of using and combining information across multiple modalities improves throughout childhood, a developmental trajectory that can be explained by general statistical learning mechanisms that are common to speech development more broadly (Getz, Nordeen, Vrabic, & Toscano, 2017).

To study AV speech integration, researchers often turn to the McGurk effect, which is an illusion created by the presentation of mismatching auditory and visual speech signals. In the original (McGurk & MacDonald, 1976) experiment, listening to the spoken syllable /bɑ/ while simultaneously watching the visual movements for /gɑ/ resulted in the illusory *fusion* perception of /dɑ/ in 98% of adult participants, thus "information from the two modalities is transformed into something new with an element not presented in either modality" (McGurk & MacDonald, 1976 p. 747). The authors took this as evidence for an important role of perceived lip movements in the perception of speech, and they argued that the fusion response was the result of averaging the common information regarding place of articulation across both modalities. Specifically, the authors conclude:

> "In a ba-voice/ga-lips presentation there is visual information for [ga] and [da] and auditory information with features common to [da] and [ba]. By responding to the common information in both modalities, a subject would arrive at the unifying percept [da]."

This explanation makes logical sense, given that the visual modality provides subjects with direct information about articulation, and it is consistent with theories of speech perception that argue for an articulatory basis for phonetic representations (Liberman & Mattingly, 1985; Fowler, 1984; Viswanathan, Magnuson, & Fowler, 2010).

Since its discovery, the original (McGurk & MacDonald, 1976) article has been cited over 7200 times (via Google Scholar citations as of December 2020), and a number of studies argue that the McGurk effect is a robust illusion, as it works when the auditory and visual signals are misaligned temporally by several hundred milliseconds (Munhall, Gribble, Sacco, & Ward, 1996; Soto-Faraco & Alsius, 2009), when the gender of the faces and voices are misaligned (Green, Kuhl, Meltzoff, & Stevens, 1991), and when the face is replaced with point-light displays (Rosenblum & Saldana, 1996). However, much of the subsequent work on this illusion has more flexibly defined what constitutes evidence for "fusion", including instances where visual information overrides the auditory component (Rosenblum & Saldana, 1992; Sams, Manninen, Surakka, Helin, & Katto, 1998)or all instances where the response deviates from the auditory component (Munhall et al., 1996; Jordan, Mccotter, & Thomas, 2000; Wilson, Alsius, Pare, & Munhall, 2016). This expanded definition differs from the original one in important ways, as it may include responses that are not fusions based on averaging of articulatory features (e.g. /θ/; Basu Mallick, Magnotti, & Beauchamp, 2015).

Factors influencing the McGurk effect

Models attempting to uncover the theoretical basis of the McGurk effect have also moved beyond the original explanation involving averaging of articulatory representations to explanations instead relying on predictive coding (Olasagasti, Bouton, & Giraud, 2015), causal inference (Magnotti & Beauchamp, 2015, 2017; Magnotti, Dzeda, Wegner-Clemens, Rennig, & Beauchamp, 2020), or more general cue integration (Massaro & Cohen, 2000; Ma, Zhou, Ross, Foxe, & Parra, 2009). These more recent models take into account the results from a variety of behavioral studies showing that the proportion of fusion responses varies depending on the stimuli used, the demands of the task, and participant characteristics. For example, Jiang and Bernstein (2011) found that physical characteristics of the stimuli can explain a large proportion of differences in listeners' responses. In addition, Basu Mallick et al. (2015) found more fusion responses (defined as /d/ or /θ/ responses) for forced-choice than open-ended designs and found the overall percentage of fusion responses was dependent on the exact stimuli used. Individuals also vary in their susceptibility to the McGurk effect, with some individuals being more likely to make fusion responses than others (Strand, Cooperman, Rowe, & Simenstad, 2014; Basu Mallick et al., 2015; Brown et al., 2018; Schwartz, 2010).

Susceptibility has been shown to differ by age, native language, and a number of clinical conditions. Fusion responses are less common in children compared to adults (McGurk & MacDonald, 1976; Tremblay et al., 2007; Hirst et al., 2018), reinforcing the notion that visual influence on speech perception increases with age (Massaro, 1984; Massaro et al., 1986; Sekiyama & Burnham, 2004; Sekiyama, 2008). Conflicting evidence exists regarding the developmental trajectory into older adulthood; for example, one study found more fusion responses in young adults ages 15 to 18 than in adults ages 27 to 58 (Pearl et al., 2009), another study found greater visual influence in older adults ages 60 to 65 than younger adults ages 19 to 21 (Sekiyama, Soshi, & Sakamoto, 2014), and yet another study found no differences in fusion rates between adults ages 18 to 35 and those ages 65 to 74 (Cienkowski & Carney, 2002). There is also conflicting evidence about the likelihood of the McGurk illusion occurring in infants, with some studies reporting effects similar to adults (Rosenblum, Schmuckler, & Johnson, 1997; Burnham & Dodd, 2004) and others showing the magnitude of the effect to be dependent on the stimuli and task (Desjardins & Werker, 2004; Tomalski, 2015).

There may also be cross-linguistic differences in the occurrence of the McGurk effect. Some research suggests a weaker McGurk effect for Japanese (Sekiyama, 2008) and Chinese (Sekiyama, 1997) listeners than for English listeners, whereas other studies have found similar occurrence of the McGurk effect between Chinese- and English-speaking participants (Chen & Hazan, 2009; Magnotti et al., 2015). Robust fusion effects have also been found with Italian (Bovo, Ciorba, Prosser, & Martini, 2009) and Finnish (Sams et al., 1998) listeners. A difference was also reported between bilingual and monolingual participants, with Korean-English bilinguals more likely to make fusion responses than monolingual English listeners (Marian, Hayakawa, Lam, & Schroeder, 2018).

Lower fusion rates have also been reported with cochlear implant users (Schorr, Fox, van Wassenhove, & Knudsen, 2005) and individuals with a number of clinical diagnoses, including specific language impairment (Norrix, Plante, & Vance, 2006; Norrix, Plante, Vance, & Boliek, 2007), autism (Bebko, Schroeder, & Weiss, 2014), Alzheimer's disease (Delbeuck, Collette, & Van der Linden, 2007), and schizophrenia (Pearl et al., 2009).

In addition to subject and task differences influencing fusion responding, the overall percentage of fusion responses reported varies greatly across studies. For example, Munhall et al. (1996) report between 1.3 and 10.7% /d/ responses across delay conditions in Experiment 1 and between 35.9 and 48.2% /d/ responses across speaking rate conditions in Experiment 2. Similarly, Green et al. (1991) found between 6 and 42% /d/ responses across their two experiments. Even MacDonald and McGurk (1978) report a lower percentage of fusion responses in their replication of the original finding, with only 64% /d/ responses in their follow-up study.

Finally, the proportion of fusion responses has been shown to be influenced by top-down cognitive processes such as expectation (Tuomainen, Andersen, Tiippana, & Sams, 2005), awareness (Palmer & Ramsey, 2012), attention (Navarra, Alsius, Soto-Faraco, & Spence, 2010), and mental imagery (Berger & Ehrsson, 2013). These results demonstrate that the McGurk effect is driven not only by bottom-up information from the stimuli, but also by top-down information, suggesting it may not simply be a low-level perceptual illusion.

Summary and goals

In summary, previous work suggests that there are a number of contextual factors influencing the occurrence of the McGurk effect. Such individual and task differences are useful because they allow us to constrain existing theories as to the basis for the effect (similar to the suggestion of Vogel & Awh, 2008 in the field of visual working memory). In the current study, we sought to replicate and extend previous work focusing on contextual variations in the McGurk effect by examining the how factors of task (forced-choice vs. open-ended), experimental stimuli (including /d/ trials vs. not), and data collection environment (lab vs. online) influence the percentage of fusion responses.

In each of our three experiments (see Table 1 for details), we compare the proportion of /d/ responses across a number of AV congruent and incongruent trial types in order to determine the strength of the McGurk effect. We first analyze the pattern of responses on the open-ended task. Specifically, we compare the proportion of auditory responses in the congruent /b/ condition to the proportion of auditory responses in the auditory /b/ - visual /g/ (i.e., McGurk) condition as a measure of overall visual influence. Then within the McGurk condition, we look at which of the non-auditory responses specifically provide evidence for "fusion" (i.e., a /d/ response). Finally, we compare the results from the open-ended task to a three-alternative forced-choice (3AFC) task where the only response options were /b/, /d/, and /g/ to determine whether the proportion of fusion responses is influenced by task.

After presenting the results of the individual experiments, we then combine all of the data to look at differences in fusion likelihood as a function of task, stimuli, and testing environment. Finding differences in fusion likelihood based on methodological changes will allow us to better determine a theoretical basis for the McGurk effect. Specifically, if the McGurk illusion is truly a perceptual effect, then individual and task differences should have little influence on the occurrence of /d/ (fusion) responses because such responses are argued to be based on the perceptual averaging and integration of auditory and visual information. However, finding that the rate of fusion responses changes based on design features would be evidence that top-down factors and context indeed influence the illusion, showing that it is the result of processes that occur after an early perceptual stage. Thus, our goal is to better situate the

Table 1 Experiment details

Experiment	Task	Stimuli	Environment
1	open-ended vs. 3AFC	/d/ included	Lab
2	open-ended vs. 3AFC	/d/ included	MTurk
3	open-ended vs. 3AFC	no /d/	MTurk

explanation for the McGurk effect within current debates in psycholinguistics (e.g., Getz & Toscano, 2019), cross-modal perception (e.g., Getz & Kubovy, 2018), and cognition more broadly (e.g., Firestone & Scholl, 2016 and subsequent commentaries) as to the extent to which top-down effects such as task demands, emotions, intentions, and linguistic representations directly influence perception. This theoretical basis will allow for a richer comparison between the McGurk illusion and AV speech integration more generally.

Experiment 1—Laboratory data collection

Experiment 1 aimed to replicate the basic McGurk effect in a lab setting using naturally produced speech. Using stimuli from two male and two female speakers (Nath & Beauchamp, 2012), we collected data using open-ended and 3AFC tasks in order to assess the prevalence of the McGurk effect (i.e., fusion responses for auditory-/b/ and visual-/g/ stimuli).

Method

Participants

Participants were Villanova students who provided informed consent and were compensated with course credit in accordance with Villanova University IRB protocols. We excluded any participants who did not report normal or corrected-to-normal vision, normal hearing, or who were not native English speakers. The final sample consisted of 86 participants (mean age = 19.0 years; 24 men).

Design

The experiment was a 2 (task; open-ended vs. 3AFC) × 6 (stimulus; aB-vB, aD-vD, aG-vG, aG-vB, aB-vD, aB-vG) mixed design, with stimulus as a within-subject factor and task as a between-subject factor (see Table 2 for AV stimulus notation and additional details). The experiment began with 12 congruent practice trials (aB-vB, aD-vD, and aG-vG from each of the four speakers) to familiarize participants with the task. The main experiment consisted of ten blocks of 24 AV trials presented in a random order, for a total of 240 trials. This means that participants completed each of the six AV trial types 40 times. Participants then completed two blocks of 12 auditory-only (AO) trials and two blocks of 12 visual-only (VO) trials at the end of the main task in order to obtain baseline responses to the auditory cues and visual cues separately. The experiment was completed in a single session lasting approximately 30 min.

Stimuli

We created our stimuli by starting with congruent AV examples of /b/, /d/, and /g/ from two male and two female speakers from the stimuli used in Nath & Beauchamp (2012; see also Basu Mallick et al., 2015). All stimuli were produced with a following /ɑ/ vowel.

The incongruent stimuli were created using iMovie version 10.1.6. We combined auditory /b/ with visual /g/ (i.e., aB-vG), auditory /b/ with visual /d/ (aB-vD), and auditory /g/ with visual /b/ (aG-vB). A summary of the AV stimuli is included in Table 2. In order to successfully line up the audio with the lip movements for each example, the auditory and visual parts of the original AV files were first separated. We then overlaid the necessary audio track to ensure that the onset of the consonant bursts matched (cf. Munhall et al., 1996; Strand et al., 2014) before removing the original audio track. From these AV stimuli, we also separated the audio and video to use in AO and VO trials, respectively.

Procedure

Participants completed the task seated in front of a computer in a sound-attenuated testing room, with stimuli presented over Sennheiser HD-558 headphones at their most comfortable level. The experimental interface was created using OpenSesame version 3.1 (Mathôt, Schreij, & Theeuwes, 2012). At the start of each block, participants were given the instructions: "After each video, select what the person said". Each trial began with a 500-ms fixation

Table 2 Audiovisual stimuli used in each experiment

Stiulus	Congruent	Experiments	Expected percept
auditory /b/–visual /b/ (aB-vB)	Yes	1, 2, 3	/b/
auditory /g/–visual /B/ (aG-vB)	No	1, 2, 3	/b/ and /g/ combined ('bga')
auditory /d/–visual /d/ (aD-vD)	Yes	1, 2	/d/
auditory /b/–visual /d/ (aB-vD)	No	1, 2	no specific prediction
auditory /g/–visual /g/ (aG-vG)	Yes	1, 2, 3	/g/
auditory /b/–visual /g/ (aB-vG)	No	1, 2, 3	/d/ (McGurk Effect)

screen, after which one of the videos would play. A response screen then appeared: half of the participants ($n = 43$) were asked to make an open-ended response by typing what the speaker said after each video and the other half ($n = 43$) made a 3AFC response (B, D, G) using a Cedrus Model RB-840 response pad to indicate their perception of what the speaker said. There was a 700-ms inter-trial interval before the next trial began.

Data analysis

R (R Core Team, 2019) was used for all analyses in the experiments reported here. Open-ended responses were divided into five mutually exclusive categories: /b/ (*e.g.* 'ba', 'bah', 'bo'), /d/ (e.g., 'da', 'dah', 'dot'), /g/ (e.g., 'ga', 'gah', 'gea'), combination (e.g., 'bdah', 'bga'), and other (e.g., 'la', 'ya', 'ma', 'na', 'tha').

Results

Overall proportions of /b/, /d/, /g/, combination (e.g., 'bga'), and other responses across all single modality and AV conditions in each experiment are presented in the Appendix (Table 3). Response proportions in Experiment 1 for the six AV conditions are displayed in Fig. 1.

Beginning with the open-ended task (Fig. 1a), participants generally responded with one of the three consonants (/b,d,g/) present in the stimuli and made only a small number of "other" or "combo" responses. On congruent trials, responses were consistent with the expected percept (/b/: 96.9%; /d/: 96.1%; /g/:92.6%). On the McGurk (aB-vG) trials, subjects rarely responded with the expected (fusion) percept, with only 10.8% /d/ responses. Instead, they primarily perceived the stimulus as consistent with the auditory component (71.9% /b/ responses) and to a lesser extent the visual component (11.0% /g/ responses). Thus, true fusion responses made up only around half of the total non-auditory responses. The other incongruent conditions produced mostly auditory-based responses (aB-vD: 70.8% /b/ responses; aG-vB: 73.0% /g/ responses).

In the 3AFC task (Fig. 1b), participants made responses consistent with the expected percept on congruent trials (/b/: 98.4%; /d/: 95.1%, /g/: 95.0%), and again primarily made responses consistent with the auditory stimulus on McGurk trials (77.1% /b/ responses). The non-auditory McGurk trial responses were approximately evenly split between visual (/g/) responses (9.7% of aB-vG total) and fusion (/d/) responses (12.3% of aB-vG total). The other incongruent trial types yielded mostly responses consistent with the auditory stimulus (aB-vD: 77.4% /b/ responses; aG-vB: 80.8% /g/ responses).

To determine whether the task (open-ended vs. 3AFC) influenced the likelihood of observing the McGurk effect, we created a logistic mixed-effect model[1] examining the proportion of fusion responses on the McGurk (aB-vG) trials, with task as a fixed effect and subject as a random effect. Task was a centered predictor, coded with 3AFC as 1 and open-ended as 0; thus, a significant positive effect would mean there were more /d/ responses in the forced-choice task compared to open-ended task. However, we did not find a main effect of task ($b = -0.02$, $SE = 0.095$, $z = -0.171$, $p = 0.864$), indicating that the number of fusion responses was similar between the open-ended and 3AFC tasks for the McGurk trials.

Discussion

Overall, the results of Experiment 1 demonstrate that, while subjects accurately perceived the congruent stimuli, the incongruent aB-vG stimuli did not produce a robust McGurk effect for either task. Because of the high proportion of auditory (/b/) responses and low proportion of fusion (/d/) responses in Experiment 1, we wanted to verify that participants were in fact watching the videos in addition to listening to the sounds. In a separate 3AFC experiment ($N = 22$), we integrated the AO and VO blocks into the main experiment so that participants had to attend to both modalities, allowing us to determine whether accuracy on the VO trials was similar to the blocked experiment. Critically, in the VO conditions, participants were 93.8, 75.6, and 47.7% accurate for /b/, /d/, and /g/ trials, respectively. These percentages show that listeners were above-chance (i.e., 33%) at identifying the correct response in the VO trials ($p < 0.001$ in each case), confirming that they were attending to the visual stimuli.

Why did the stimuli not produce a McGurk effect in Experiment 1? One possibility is that the undergraduate subject population differed from other groups of subjects in previous experiments investigating the effect. One factor in particular that we were interested in is the effect of age. Given previous studies showing age differences in the proportion of fusion effects (McGurk & MacDonald, 1976; Tremblay et al., 2007; Pearl et al., 2009; Sekiyama et al., 2014; Cienkowski & Carney, 2002), we next asked whether the McGurk effect occurs more often in a population with a wider range of ages.

[1] All logistic mixed-effect analyses were run using the R (R Core Team, 2019) package lme4 (Bates, Maechler, & Bolker, 2014) and reported p values were calculated based on the Wald z-statistics for each parameter in the model summary.

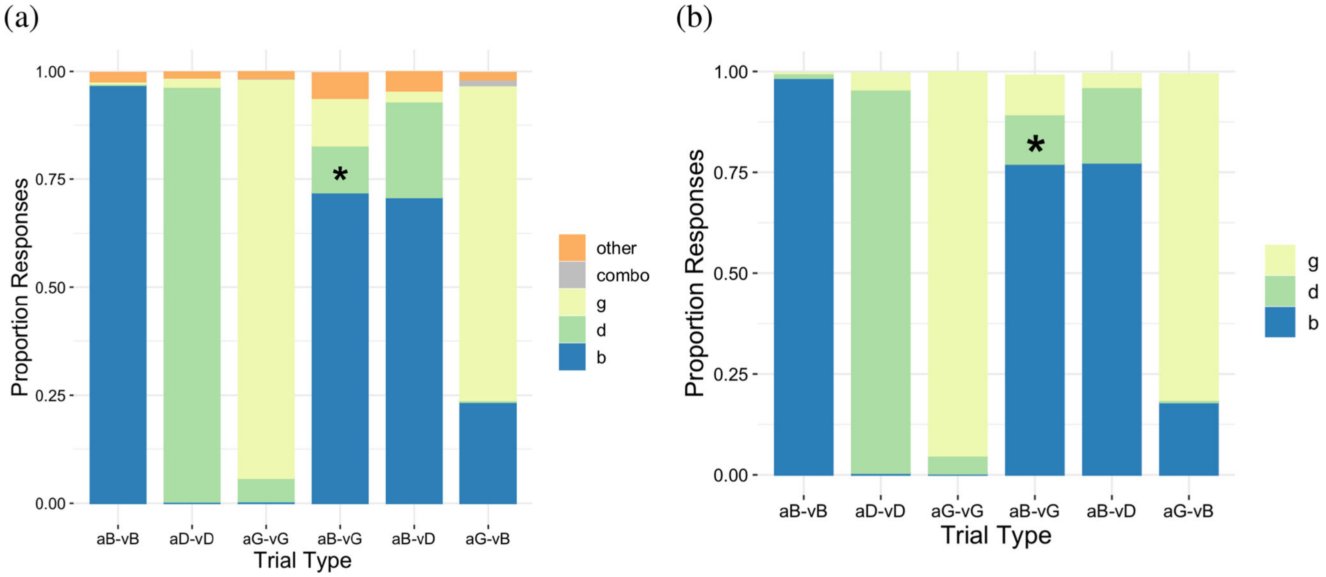

Fig. 1 Proportion of responses by trial type for Experiment 1. **a** Results of the open-ended task. **b** Results of the 3AFC task. McGurk fusion responses (i.e., /d/ responses for aB-vG trials) are marked with a * for easy comparison across experiments

Experiment 2—Online data collection

We investigated whether the testing environment and subject characteristics play a role in the occurrence of the McGurk illusion by using Amazon.com's Mechanical Turk (MTurk) service to conduct the experiment online. This provided an opportunity to test the McGurk effect in a broader participant population, particularly with a broader age range of subjects. We again collected data using both an open-ended and 3AFC task.

Method

Participants

We used MTurk to recruit 76 participants (36 men; mean age = 37.6, age range, 21-72 years) to complete the experiment. Participants in this experiment provided informed consent and were compensated $3.63 for participating in the 30-minute experiment in accordance with Villanova IRB protocols. All participants reported normal or corrected-to-normal vision, normal hearing, and were native English speakers.

Design

The experiment was a 2 (task) × 6 (stimulus) mixed design, using the same tasks and stimuli as Experiment 1. The experiment began with 12 AV congruent practice trials. The main task consisted of eight blocks of 24 AV trials presented in a random order, for a total of 192 trials. Participants also completed two blocks of 12 AO and two blocks of 12 VO trials. The experiment took approximately 30 min to complete.

Procedure

Before the beginning of the experiment, we checked that participants could hear the auditory stimuli using a procedure similar to Toscano and Lansing (2019). First, subjects were asked to wear headphones during the experiment and to indicate via a text box the brand of headphones they were wearing. These were reviewed after the experiment to ensure that the subjects provided reasonable answers. Second, subjects were presented with a 1-kHz calibration tone and asked to adjust their computer volume to a comfortable level. Finally, subjects were given three trials in which they heard individual words and had to choose what they heard from a list of six alternatives (i.e., ball, cube, dots, girl, pear, tame). Subjects had to answer correctly on each of these trials to begin the main experiment. This was done as a final check to ensure that they could not only hear the stimuli, but also identify them accurately.

Subjects then began the main experiment. Throughout the experiment, the videos appeared on the screen one at a time, with the instructions "What did the speaker say?" underneath. Approximately half of the participants ($n = 39$) were presented with a blank text box to provide an open-ended response and the other half ($n = 37$) were asked to select buttons corresponding to one of three choices (Ba, Da, Ga).

Results

The proportions of responses for the six AV conditions are displayed in Fig. 2. As in Experiment 1, in the open-ended task (Fig. 2a), participants correctly recognized the stimuli on the congruent trials (/b/: 92.0%; /d/: 95.3%; /g/: 95.4%). On McGurk (aB-vG) trials, subjects made 37.0% auditory-based (i.e., /b/) responses, 5.7% visual-based (i.e., /g/) responses, 16.7% fusion (/d/) responses, and 40.4% "other" responses. Among the "other" responses, the modal response was /θ/, with /p/ and /l/ as other common responses. Thus, true fusion responses were much less common than other non-/b/ responses. The aB-vD condition yielded a pattern of responses similar to the McGurk (aB-vG) condition. The aG-vB condition produced primarily auditory-based (i.e., /g/) responses (76.5%).

In the 3AFC task (Fig. 2b), participants were again highly accurate in the congruent conditions (/b/: 93.8%; /d/: 93.2%; /g/: 94.5%). On the McGurk trials, participants made 50.8% auditory-based (/b/) responses, 42.6% fusion (/d/) responses, and 6.6% visual-based (/g/) responses. Thus, although fusion responses were not the modal response type, the proportion of fusion responses was higher than in the open-ended task. The aB-vD condition again showed a pattern of responses similar to the McGurk condition, and the aG-vB condition produced primarily auditory-based (/g/) responses (95.1%).

As in Experiment 1, we created a logistic mixed-effect model to compare the proportion of fusion responses on the McGurk trials between the open-ended and 3AFC tasks. In addition to the fixed effect of task, we included age and the interaction between task and age in the model. We found a main effect of task ($b = 0.49$, $SE = 0.118$, $z = 4.14$, $p < 0.001$), such that there were significantly more fusion responses in the 3AFC than in the open-ended task. There was also a main effect of age ($b = 0.01$, $SE = 0.005$, $z = 2.03$, $p = 0.042$), with older participants making more fusion responses. There was no interaction between task and age.

Discussion

Compared with Experiment 1, the current experiment produced a stronger McGurk effect, though auditory-based responses were still more common than fusion responses. The difference between the two experiments may partly be attributable to an age effect, as older participants in the current experiment were more likely to show the effect.

The results also reveal a task effect, with more fusion responses in the 3AFC condition. This suggests that the response set plays a role in whether or not subjects experience the illusion. Similarly, the *stimulus* set may also play a role. For instance, previous work has demonstrated that the range of stimulus values along an acoustic continuum affects listeners' responses in speech categorization experiments (Brady & Darwin, 1978; Rosen, 1979). A similar effect could occur in experiments involving the McGurk effect. Thus, by eliminating the /d/ stimuli, we might observe a further increase in the incidence of the illusion. This possibility was addressed in Experiment 3.

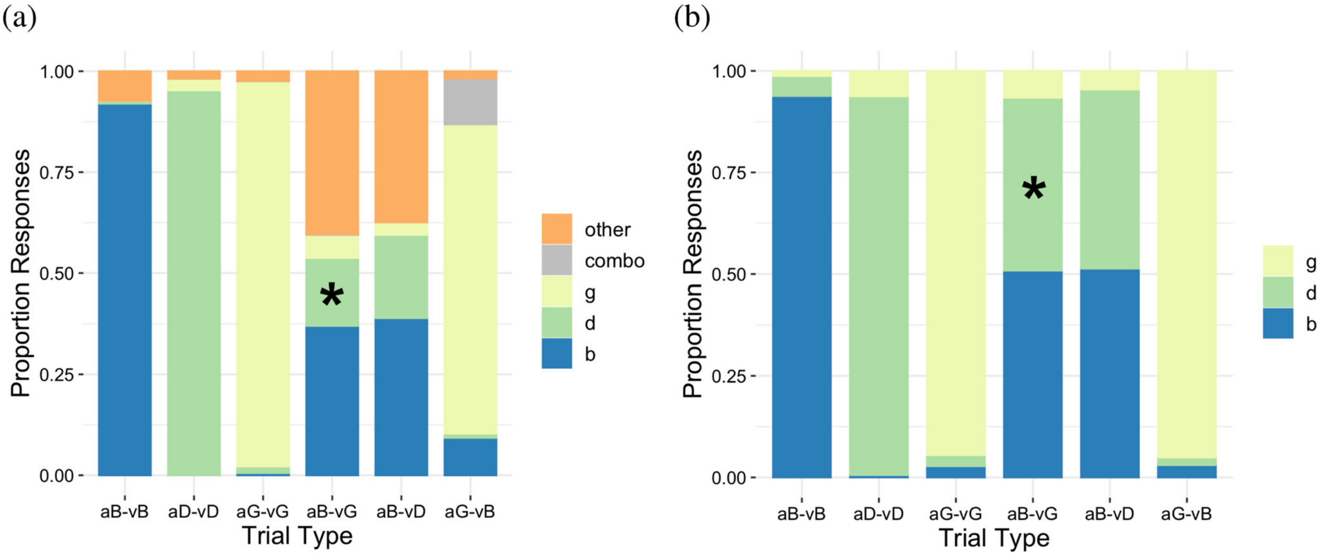

Fig. 2 Proportion of responses by trial type for Experiment 2. **a** Results of the open-ended task. **b** Results of the 3AFC task. McGurk fusion responses (i.e., /d/ responses for aB-vG trials) are marked with a * for easy comparison across experiments

Experiment 3—Different stimulus set

Participants performed the same two tasks as the previous experiments (3AFC and open-ended), but we removed the congruent /d/ condition and the aB-vD condition. As a result, there are no conditions in the experiment that include the production of the phoneme /d/; the McGurk condition is the "closest match" to a /d/ stimulus. The four remaining stimuli (aB-vB, aG-vG, aB-vG, and aG-vB) match the stimulus set included in the original McGurk effect (McGurk & MacDonald, 1976) and more recent work examining individual differences in the effect likelihood (Basu Mallick et al., 2015). If this change affects the likelihood of the McGurk effect occurring, we would expect more overall fusion responses in the aB-vG condition here than in the previous experiments.

Method

Participants

We used MTurk to recruit 76 participants (35 men; mean age = 41.2, age range, 25-65) to complete the experiment. All participants reported normal or corrected-to-normal vision, normal hearing, and were native English speakers. Experiment 3 was approved by the University of San Diego Institutional Review Board, and subjects provided informed consent and received $3.63 for their participation.

Design

The experiment was a 2 (task; 3AFC vs. open-ended) × 4 (stimulus; aB-vB, aG-vG, aB-vG, aG-vB) mixed design. The experiment began with eight AV congruent practice trials (aB-vB and aG-vG from each of the four speakers). The main task consisted of ten blocks of 16 AV trials presented in a random order, for a total of 160 trials. Participants also completed two blocks of eight AO and two blocks of eight VO trials. The experiment took approximately 30 min to complete.

Procedure

The procedure was identical to Experiment 2, with approximately half of the participants ($n = 39$) completing an open-ended task, and the other half of the participants ($n = 37$) completing a 3AFC task.

Results

The proportion of responses for the six AV conditions are displayed in Fig 3. In the open-ended task (Fig. 3a), participants were accurate in both congruent conditions (/b/: 94.9% correct; /g/: 97.8% correct). The McGurk (aB-VG) condition led to 45.4% auditory (/b/) responses, 19.3% fusion (/d/) responses, 8.4% visual (/g/) responses, and 26.9% "other" responses. Among the "other" responses, the

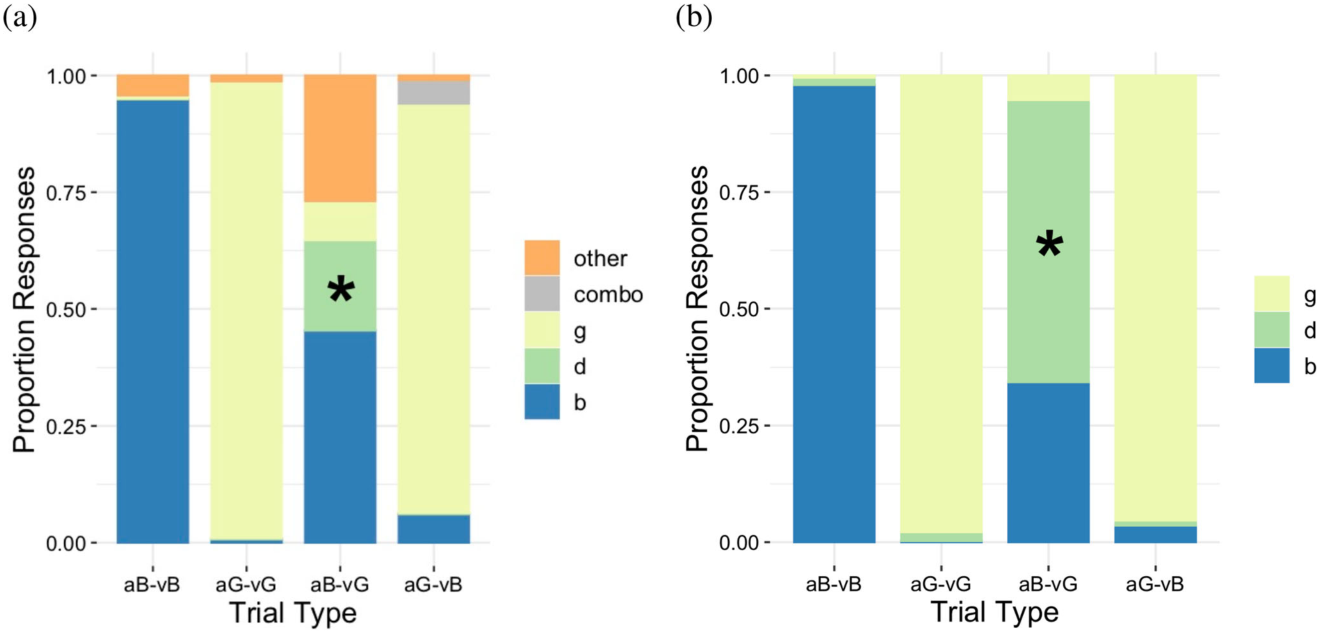

Fig. 3 Proportion of responses by trial type for Experiment 3. **a** Results of the open-ended task. **b** Results of the 3AFC task. McGurk fusion responses (i.e., /d/ responses for aB-vG trials) are marked with a * for easy comparison across experiments

modal response was /θ/ followed closely by /l/. The aG-vB condition yielded primarily auditory-based responses (87.6% /g/).

In the 3AFC task (Fig. 3b), participants were again highly accurate in the two congruent conditions (/b/: 98.0% correct; /g/: 97.8% correct). On McGurk trials, subjects made 34.3% auditory-based responses, 60.4% fusion responses, and 5.3% visual-based responses. This is the only McGurk condition across all three experiments where the fusion (/d/) response was the modal response. The aG-vB condition produced mostly auditory-based responses (95.3% /g/).

As in the previous experiments, we created a logistic mixed-effects model to compare the proportion of fusion responses on the McGurk trials between the Experiment 3 open-ended and 3AFC tasks, with task, age, and their interaction as fixed effects. As in Experiment 2, we found a main effect of task ($b = 2.17$, $SE = 0.451$, $z = 6.19$, $p < 0.001$), such that there were significantly more fusion responses with a 3AFC than open-ended task, and a main effect of age ($b = 0.04$, $SE = 0.018$, $z = 2.52$, $p = 0.012$), with older participants showing more fusion responses. The interaction was not significant.

Discussion

Experiment 3 produced the strongest McGurk effect among the overall set of experiments. This was driven by the removal of the congruent and incongruent /d/ conditions, which resulted in no trials where /d/ was the correct response, and consequently, a greater proportion of /d/ responses on the McGurk trials. Thus, these results suggest that the McGurk effect is driven by the stimulus set and response options given to subjects, in addition to individual differences and variation in the effectiveness of specific McGurk stimuli.

Subject, stimulus set, and task differences across Experiments 1-3

Given that previous work has shown individual differences in McGurk susceptibility (Strand et al., 2014; Basu Mallick et al., 2015; Brown et al., 2018), we wanted to explore how the proportion of /d/ responses on the McGurk trials (aB-vG) varied across participants in the experiments. To do this, we combined all 199 participants into a single analysis. We found that the overall percentage of fusion responses in the McGurk condition was 20.1% across experiments, which is much lower than the original experiments (McGurk & MacDonald, 1976; MacDonald & McGurk, 1978), though more in line with more recent studies (e.g., Munhall et al., 1996; Green et al., 1991; Basu Mallick et al., 2015). Additionally, this proportion varied greatly across participants, with a range from 0% to 100% /d/ responses on the McGurk trials (Fig. 4). Further, most participants showed almost no fusion responses at all.

Next, to compare how differences in task (forced-choice vs. open-ended), stimuli (including /d/ vs. not), and environment (lab vs. online) affected the proportion of fusion responses on the McGurk trials, we created a logistic mixed-effects model with fixed effects of task, stimuli, environment and all possible interactions with the between-experiment factors: task × stimuli and task × environment.[2] The model also included subject as a random effect.

All fixed effect factors were numerically coded, centered predictors. *Task* was coded with 3AFC as 1 and open-ended as 0; thus, a significant positive effect would mean more fusion responses on forced-choice tasks. *Stimulus* was coded with experiments that had no /d/ stimuli as 1 and experiments including /d/ as 0; thus, a significant positive effect would mean more fusion responses in Experiment 3 than Experiments 1–2. *Environment* was coded with the lab as 1 and MTurk as 0; thus, a significant positive effect would mean more fusion responses in Experiment 1 than Experiments 2–3.

We found a main effect of *task* ($b = 1.63$, $SE = 0.369$, $z = 4.43$, $p < 0.001$), such that there were more fusion responses overall with a 3AFC than open-ended task. There was also a main effect of *stimulus type* ($b = 0.72$, $SE = 0.283$, $z = 2.56$, $p = 0.010$), with more fusion responses overall when /d/ stimuli were not present in the experiment. We also found a main effect of *environment* ($b = -2.48$, $SE = 0.425$, $z = -5.84$, $p < 0.001$), such that there were more fusion responses overall on MTurk than in the lab.

Several interactions were also significant. We found an interaction between task and environment ($b = -1.75$, $SE = 0.796$, $z = -2.20$, $p = 0.028$), such that there was a difference in the proportion of fusion responses between 3AFC and open-ended on MTurk ($b = 2.50$, $SE = 0.408$, $z = 6.12$, $p < 0.001$), but no difference in the lab ($b = 0.30$, $p = 0.696$). We also found an interaction between task and stimuli ($b = 1.22$, $SE = 0.57$, $z = 2.12$, $p = 0.034$), such that there was a difference in the proportion of fusion responses between experiments including /d/ stimuli and experiments excluding /d/ stimuli on the 3AFC task ($b = 1.35$, $SE = 0.58$, $z = 2.33$, $p = 0.020$), but no difference on the open-ended task ($b = 0.13$, $p = 0.269$).

Finally, we conducted an analysis that included age as a fixed effect predictor in the logistic mixed-effect model to determine whether the differences in fusion proportions across task, stimuli, and environment would remain when

[2] Because this analysis did not have a fully crossed design between stimuli and environment, this is the maximal interaction structure for fixed effects.

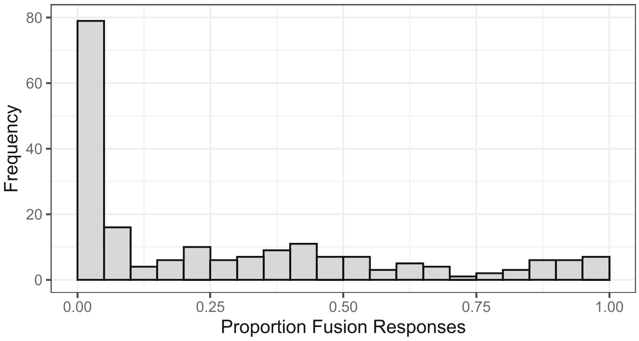

Fig. 4 Histogram showing the distribution of fusion responses in the McGurk condition (aB-vG) across all participants in Experiments 1-3

taking participant age into account. We found a main effect of age ($b = 0.10$, $SE = 0.008$, $z = 12.30$, $p < 0.001$) and task ($b = 1.76$, $SE = 0.38$, $z = 4.61$, $p < 0.001$), but no main effects of stimuli ($b = 0.25$, $p = 0.394$) or environment ($b = -0.58$, $p = 0.216$). The interactions between task and environment ($b = -1.94$, $SE = 0.834$, $z = -2.33$, $p = 0.020$) and task and stimuli ($b = 1.21$, $SE = 0.592$, $z = 2.04$, $p = 0.042$) also remained significant.

General discussion

The results of these three experiments demonstrate that several factors (task, stimulus set, and participant characteristics) have a large influence the occurrence of the McGurk effect. Across experiments, using the same stimuli to elicit the McGurk effect, we found fusion responses ranging from 10.8 to 60.4%. Thus, there is a high degree of variability in the likelihood of experiencing the McGurk effect across factors that are unrelated to the perceptual information provided by the stimulus.

Additionally, the number of fusion responses varied greatly by individual, similar to previous studies (Strand et al., 2014; Basu Mallick et al., 2015; Brown et al., 2018; Schwartz, 2010), while the majority of our participants across experiments actually showed *no* fusion responses whatsoever. One relevant participant characteristic that predicts differences in fusion is age, with older listeners more likely to show fusion responses (McGurk & MacDonald, 1976; Tremblay et al., 2007; Sekiyama et al., 2014). We replicated this finding in Experiments 2 and 3, highlighting one potential factor influencing McGurk susceptibility. This is particularly noteworthy because other research has shown that McGurk susceptibility does *not* relate to individual differences in attentional control, processing speed, working memory capacity, or auditory perceptual gradiency (Brown et al., 2018). Though an understanding of why age influences the likelihood of experiencing the McGurk effect is better left to future research, this result is in line with previous work suggesting that phonetic cue weights continue to change across the lifespan (even in the absence of changes related to hearing loss; Toscano & Lansing, 2019).

Further, the specific demands of the task influenced the likelihood of fusion responses in the current experiments. First, in each experiment, we found more fusion responses in the 3AFC task than in the open-ended task (as in Basu Mallick et al. 2015). Second, we found more fusion responses on MTurk than in the lab, suggesting possible roles for attention (Navarra et al., 2010), age, or other factors in McGurk susceptibility. This result differs from Magnotti et al. (2018), who found similar results for in-person and online experiments, though there are a number of differences between the two studies that might explain the different pattern of results (e.g., number of repetitions of the stimulus within a trial, number of trials, which types of responses count as fusion responses). Third, we found more fusion responses when no true /d/ examples were included as stimuli. This suggests that adding a congruent /d/ may provide a reference of how that speaker produces the /d/ syllable, making participants less susceptible to the illusion. Critically, this effect is not due to perceptual differences in the McGurk stimuli, but rather, relates to the subject's knowledge of how a particular talker produces speech sounds. Such an effect is inconsistent with a low-level perceptual interpretation of the McGurk effect.

Given the considerable variability in the occurrence of fusion responses across tasks, stimulus sets, and testing environment, we argue that fusion responses are more likely the result of decision-level processes than early perceptual ones. For example, when only given three options in an incongruent AV setting, listeners may be more likely to choose the response option that does not align with the auditory or visual stimulus. Conversely, when unconstrained in their options, listeners can more accurately report what they perceived rather than choosing what they must have perceived from the given options. In this case, the source of their response is not low-level perceptual interactions between the auditory and visual stimuli.

Our experiments have a similar limitation as most studies of the McGurk effect, which almost always use isolated syllables rather than real words (though see Dekle, Fowler, & Funnell, 1992; Sams et al., 1998; Windmann, 2004 for exceptions). Using isolated syllables is another way that this illusion is removed from everyday speech perception settings. Investigating the likelihood of McGurk fusion responses in words would be another way to look at how familiarity, memory, and context effects influence the illusion, showing the importance of top-down contextual influences rather than perceptual processes alone in explaining the effect.

The inclusion of top-down influences is in line with a number of recent models of the McGurk effect that include parameters for participants and stimuli and allow for top-down predictions across modalities (Magnotti & Beauchamp, 2017; Ma et al., 2009; Olasagasti et al., 2015). We agree that incorporating such factors into models of AV speech perception is critical. Further, we argue explicitly that such individual differences and top-down influences necessarily mean the McGurk illusion is not a *perceptual* effect, and rather, fusion responses happen at a higher cognitive level. This conclusion is also consistent with more general work on cross-modal correspondences that shows how top-down influences undermine the assumption of automaticity in multisensory integration (e.g., Getz & Kubovy, 2018).

We can also consider where the McGurk effect fits into the broader literature on AV speech perception. Two recent reviews of research on the McGurk illusion come to opposite conclusions regarding how it compares to AV speech integration more generally. Marques, Lapenta, Costa, and Boggio (2016) argue that the robustness of the McGurk effect makes in a useful tool for investigating unconscious multisensory integration processes, whereas Alsius, Pare, and Munhall (2017) highlight a number of important differences between the McGurk illusion and congruent AV speech processing, urging caution when generalizing McGurk effect results to natural AV speech integration. Further, other work shows no relationship between McGurk susceptibility and AV speech benefit (i.e., speech-in-noise processing with and without a visual signal; Van Engen, Xie, & Chandrasekaran, 2017), even though both tasks are used as measures of AV integration (though note that speech-in-noise perception is also influenced by task and subject differences; see Magnotti et al., 2020 for an alternative explanation of this weak correlation). The current results support the idea that the McGurk illusion is not a strong proxy for AV speech integration, though we reach this conclusion for different reasons than others, with Alsius and colleagues still framing differences between the McGurk effect and congruent speech in terms of "perceptual mechanisms," whereas we argue that the mounting evidence for top-down influences argues against the McGurk illusion as a perceptual effect.

If the goal is to understand audiovisual speech *perception*, designing tasks that more directly measure perception should be the focus of future research. Tasks that involve asking listeners to repeat what a speaker said and tasks that use forced-choice responses are confounded with speech production and decision-making processes. Thus, these tasks may not tell us about AV perceptual integration. Indeed, this is an issue that affected work investigating categorical perception of speech for decades: while the original experiments suggested that listeners perceive speech categorically (Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967), later work demonstrated that these effects depended on the type of discrimination task used (ABX vs. 4IAX discrimination tasks; Pisoni & Lazarus, 1974; Schouten, Gerrits, & van Hessen, 2003), the response options given to the listener (e.g., forced-choice vs. category goodness ratings; Miller, 1994), and individual differences between subjects (Kapnoula, Winn, Kong, Edwards, & Mcmurray, 2017). In a similar way, the McGurk effect, along with the stimuli and tasks used to elicit it, may not be the best test case for understanding AV integration in speech more broadly. Other creative behavioral methods as well as cognitive neuroscience methods, such as the event-related potential (ERP) technique, which more directly captures early perceptual processes (Toscano, McMurray, Dennhardt, & Luck, 2010) and can measure the influence of top-down effects on them (Getz & Toscano, 2019), may provide valuable information going forward.

This is not to say that using incongruent auditory and visual speech stimuli are irrelevant; clearly, the use of incongruent stimuli is one important way to investigate trade-offs between modalities. Instead of focusing on the very specific example of the McGurk illusion, however, we would recommend that researchers manipulate multiple acoustic and visual cue values along a continuum from congruent to incongruent (cf. Massaro, 2017; Getz et al., 2017) in order to investigate the competition between cues in a more generalizable way. It is thus not the case that all measures of audiovisual speech behavior will be subject to similar context effects as the McGurk effect; but it is necessary for researchers to come up with new paradigms to study true *perceptual* effects.

These conclusions and recommendations echo those of other researchers who have encouraged the field to move beyond the McGurk effect as a measure of audiovisual integration or multisensory binding (Alsius et al., 2017; Massaro, 2017). Here, we argue that not only is the McGurk effect a poor example of AV integration, it is also not a robust perceptual illusion. Instead, it is an effect that exists only for some participants under specific task situations and stimulus conditions. In order to better understand audiovisual speech integration, we would challenge future researchers to think carefully about the theoretical basis of the McGurk effect, how it is defined, and what it can tell us more generally about how top-down effects directly influence perception.

Appendix Table

Table 3 Proportion of /b/, /d/, /g/, combination, and other responses by experiment and trial type

Experiment	Trial type	/b/	/d/	/g/	combo	other
1: open	AO /b/	0.988	0.000	0.003	0.000	0.006
	AO /d/	0.000	0.994	0.000	0.000	0.003
	AO /g/	0.000	0.000	0.994	0.003	0.003
	VO /b/	0.956	0.003	0.009	0.003	0.029
	VO /d/	0.012	0.831	0.105	0.000	0.047
	VO /g/	0.012	0.369	0.558	0.000	0.049
	AV /b/	0.969	0.003	0.005	0.000	0.020
	AV /d/	0.003	0.961	0.021	0.000	0.013
	AV /g/	0.004	0.053	0.926	0.001	0.015
	A/b/-V/g/	0.719	0.108	0.110	0.000	0.059
	A/b/-V/d/	0.708	0.221	0.027	0.000	0.043
	A/g/-V/b/	0.234	0.004	0.730	0.013	0.016
1: 3AFC	AO /b/	0.962	0.026	0.000		
	AO /d/	0.003	0.985	0.006		
	AO /g/	0.000	0.006	0.988		
	VO /b/	0.988	0.003	0.006		
	VO /d/	0.003	0.890	0.093		
	VO /g/	0.015	0.387	0.590		
	AV /b/	0.984	0.012	0.001		
	AV /d/	0.005	0.951	0.041		
	AV /g/	0.003	0.045	0.950		
	A/b/-V/g/	0.770	0.123	0.097		
	A/b/-V/d/	0.774	0.187	0.033		
	A/g/-V/b/	0.180	0.006	0.808		
2: open	AO /b/	0.798	0.019	0.000	0.000	0.183
	AO /d/	0.000	0.942	0.032	0.000	0.026
	AO /g/	0.000	0.022	0.949	0.000	0.029
	VO /b/	0.885	0.019	0.022	0.000	0.064
	VO /d/	0.016	0.551	0.138	0.013	0.279
	VO /g/	0.013	0.343	0.429	0.013	0.196
	AV /b/	0.920	0.006	0.001	0.000	0.073
	AV /d/	0.000	0.953	0.029	0.000	0.018
	AV /g/	0.006	0.016	0.954	0.001	0.024
	A/b/-V/g/	0.370	0.167	0.057	0.002	0.404
	A/b/-V/d/	0.389	0.205	0.031	0.000	0.374
	A/g/-V/b/	0.093	0.010	0.765	0.114	0.018

Table 3 (continued)

Experiment	Trial Type	/b/	/d/	/g/	combo	other
2: 3AFC	AO /b/	0.932	0.044	0.024		
	AO /d/	0.024	0.905	0.071		
	AO /g/	0.020	0.024	0.953		
	VO /b/	0.902	0.078	0.020		
	VO /d/	0.034	0.821	0.145		
	VO /g/	0.041	0.497	0.463		
	AV /b/	0.938	0.050	0.013		
	AV /d/	0.006	0.932	0.062		
	AV /g/	0.028	0.027	0.945		
	A/b/-V/g/	0.508	0.426	0.066		
	A/b/-V/d/	0.514	0.441	0.046		
	A/g/-V/b/	0.030	0.019	0.951		
3: open	AO /b/	0.814	0.038	0.013	0.000	0.135
	AO /g/	0.012	0.001	0.952	0.006	0.022
	VO /b/	0.958	0.001	0.016	0.000	0.016
	VO /g/	0.058	0.151	0.683	0.003	0.106
	AV /b/	0.949	0.001	0.001	0.000	0.042
	AV /g/	0.007	0.003	0.978	0.000	0.012
	A/b/-V/g/	0.454	0.193	0.084	0.000	0.269
	A/g/-V/b/	0.061	0.002	0.876	0.051	0.010
3: 3AFC	AO /b/	0.899	0.091	0.010		
	AO /g/	0.000	0.030	0.970		
	VO /b/	0.956	0.010	0.034		
	VO /g/	0.020	0.456	0.524		
	AV /b/	0.980	0.016	0.004		
	AV /g/	0.003	0.019	0.978		
	A/b/-V/g/	0.343	0.604	0.053		
	A/g/-V/b/	0.036	0.011	0.953		

Acknowledgements We would like to thank Abby Benecke, Elke Nordeen, and Sarah Vrabic for discussion of these issues. We would also like to thank the research assistants in the Word Recognition & Auditory Perception Lab for assistance with data collection. LMG was supported by a Mendel Science Experience Post-doctoral Fellowship from the Villanova University College of Liberal Arts and Sciences and a Faculty Research Grant from the University of San Diego.

Open Practices Statement None of the experiments were preregistered. Data and materials for all experiments are available upon request from the corresponding author.

References

Aldridge, M., Braga, E., Walton, G., & Bower, T. (1999). The intermodal representation of speech in newborns. *Developmental Science*, *2*, 42–46.

Alsius, A., Pare, M., & Munhall, K. G. (2017). Forty years after Hearing lips and seeing voices: the McGurk effect revisited. *Multisensory Research*, *31*, 111–144.

Barutchu, A., Danaher, J., Crewther, S. G., Innes-Brown, H., Shivdasani, M. N., & Paolini, A. G. (2010). Audiovisual integration in noise by children and adults. *Journal of Experimental Child Psychology*, *105*, 38–50.

Basu Mallick, D., Magnotti, J. F., & Beauchamp, M. S. (2015). Variability and stability in the McGurk effect: Contributions of participants, stimuli, time, and response type. *Psychonomic Bulletin & Review*, *22*(5), 1299–1307.

Bates, D., Maechler, M., & Bolker, B. (2014). `lme4`: Linear mixed-effects models using Eigen and S4 [Computer software manual]. Retrieved from http://CRAN.R-project.org/package=lme4 (R package version 1.1-7).

Bebko, J. M., Schroeder, J. H., & Weiss, J.A. (2014). The McGurk effect in children with autism and Asperger syndrome. *Autism Research*, *7*, 50–59.

Berger, C. C., & Ehrsson, H. H. (2013). Mental imagery changes multisensory perception. *Current Biology*, *23*, 1367–1372.

Bovo, R., Ciorba, A., Prosser, S., & Martini, A. (2009). The McGurk phenomenon in Italian listeners. *Acta Otorhinolaryngologica Italica*, *29*(4), 203–208.

Brady, S. A., & Darwin, C. J. (1978). Range effect in the perception of voicing. *The Journal of the Acoustical Society of America*, *63*(5), 1556–1558.

Brown, V. A., Hedayati, M., Zanger, A., Mayn, S., Ray, L., & Dillman-Hasso, N. (2018). What accounts for individual differences in susceptibility to the McGurk effect? *PLoS One*, *13*(11), e0207160.

Burnham, D., & Dodd, B. (2004). Auditory-visual speech integration by prelinguistic infants: Perception of an emergent consonant in the McGurk effect. *Developmental Psychobiology*, *45*(4), 204–220.

Chen, Y., & Hazan, V. (2009). Developmental factors and the non-native speaker effect in auditory-visual speech perception. *Journal of the Acoustical Society of America*, *126*(2), 858–865.

Cienkowski, K. M., & Carney, A. E. (2002). Auditory-visual speech perception and aging. *Ear and Hearing*, *23*, 439–449.

Dekle, D. J., Fowler, C. A., & Funnell, M. G. (1992). Audiovisual integration in perception of real words. *Perception & Psychophysics*, *51*(4), 355–362.

Delbeuck, X., Collette, F., & Van der Linden, M. (2007). Is Alzheimer's disease a disconnection syndrome? Evidence from a crossmodal audio-visual illusory experiment. *Neuropsychologia*, *45*(14), 3315–3323.

Desjardins, R., & Werker, J. (2004). Is the integration of heard and seen speech mandatory for infants?. *Developmental Psychobiology*, *45*(4), 187–203.

Dodd, B. (1979). Lip reading in infants: Attention to speech presented in- and out-of-synchrony. *Cognitive Psychology*, *11*(4), 478–484.

Firestone, C., & Scholl, B. J. (2016). Cognition does not affect perception: Evaluating the evidence for "top-down" effects. *Behavioral and Brain Sciences*, *39*, e229.

Fowler, C. A. (1984). Segmentation of coarticulated speech in perception. *Perception & Psychophysics*, *36*, 359–368.

Getz, L. M., & Kubovy, M. (2018). Questioning the automaticity of audiovisual correspondences. *Cognition*, *175*, 101–108.

Getz, L. M., Nordeen, E., Vrabic, S., & Toscano, J. (2017). Modeling the development of audiovisual cue integration in speech perception. Brain Sciences, 7(3) article 32.

Getz, L. M., & Toscano, J. C. (2019). Electrophysiological evidence for top-down lexical influences on early speech perception. *Psychological Science*, *30*(6), 830–841.

Gilbert, J. L., Lansing, C. R., & Garnsey, S. M. (2012). Seeing facial motion affects auditory processing in noise. *Attention, Perception, & Psychophysics*, *74*(8), 1761–1781.

Green, K. P., Kuhl, P. K., Meltzoff, A. N., & Stevens, E. B. (1991). Integrating speech information across talkers, gender, and sensory modality: Female faces and male voices in the McGurk effect. *Perception & Psychophyics*, *50*, 524–536.

Guellaï, B., Streri, A., Chopin, A., Rider, D., & Kitamura, C. (2016). Newborns' sensitivity to the visual aspects of infant-directed speech: Evidence from point-line displays of talking faces. Journal of Experimental Psychology: Human Perception and Performance Advance online publication.

Hirst, R. J., Stacey, J. E., Cragg, L., Stacey, P. C., & Allen, H. A. (2018). The threshold for the McGurk effect in audio-visual noise decreases with development. *Scientific Reports*, *8*(1), 1–12.

Jiang, J., & Bernstein, L. E. (2011). Psychophysics of the McGurk and other audiovisual speech integration effects. *Journal of Experimental Psychology: Human Perception and Performance*, *37*(4), 1193–1209.

Jordan, T. R., Mccotter, M. V., & Thomas, S. M. (2000). Visual and audiovisual speech perception with color and gray scale facial images. *Perception & Psychophysics*, *62*, 1394–1404.

Kapnoula, E. C., Winn, M. B., Kong, E. J., Edwards, J., & Mcmurray, B. (2017). Evaluating the sources and functions of gradiency in phoneme categorization: An individual differences approach. *Journal of Experimental Psychology: Human Perception and Performance*, *43*(9), 1594–1611.

Kitamura, C., Guellaï, B., & Kim, J. (2014). Motherese by eye and ear: Infants perceive visual prosody in point-line displays of talking heads. *PLoS ONE*, *9*(10), e111467.

Knowland, V. C., Mercure, E., Karmiloff-Smith, A., Dick, F., & Thomas, M. S. (2014). Audio-visual speech perception: A developmental ERP investigation. *Developmental Science*, *17*(1), 110–124.

Kuhl, P. K., & Meltzoff, A. N. (1984). The intermodal representation of speech in infants. *Infant Behavior and Development*, *7*(3), 361–381.

Lewkowicz, D. J. (2010). Infant perception of audio-visual speech synchrony. *Developmental Psychology*, *46*(1), 66–77.

Liberman, A. M., Cooper, F. S., Shankweiler, D. P., & Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological Review*, *74*, 431–461.

Liberman, A. M., & Mattingly, I. G. (1985). The motor theory of speech perception revised. *Cognition*, *21*, 1–36.

Ma, W. J., Zhou, X., Ross, L. A., Foxe, J. J., & Parra, L. C. (2009). Lip-reading aids word recognition most in moderate noise: a Bayesian explanation using high-dimensional feature space. *PloS One*, *4*(3), e4638.

MacDonald, J., & McGurk, H. (1978). Visual influences on speech perception processes. *Attention, Perception, & Psychophysics*, *24*(3), 253–257.

Magnotti, J. F., & Beauchamp, M. S. (2015). The noisy encoding of disparity model of the McGurk effect. *Psychonomic Bulletin & Review*, *22*(3), 701–709.

Magnotti, J. F., & Beauchamp, M. S. (2017). A causal inference model explains perception of the McGurk effect and other incongruent audiovisual speech. *PLoS Computational Biology*, *13*(2), e1005229.

Magnotti, J. F., Dzeda, K. B., Wegner-Clemens, K., Rennig, J., & Beauchamp, M. S. (2020). Weak observer–level correlation and strong stimulus-level correlation between the McGurk effect and audiovisual speech-in-noise: A causal inference explanation. *Cortex*, *133*, 371–383.

Magnotti, J. F., Mallick, D. B., Feng, G., Zhou, B., Zhou, W., & Beauchamp, M. S. (2015). Similar frequency of the McGurk effect in large samples of native Mandarin Chinese and American English speakers. *Experimental Brain Research*, *233*(9), 2581–2586.

Magnotti, J. F., Smith, K. B., Salinas, M., Mays, J., Zhu, L. L., & Beauchamp, M. S. (2018). A causal inference explanation for enhancement of multisensory integration by co-articulation. *Scientific Reports*, *8*(1), 1–10.

Maidment, D. W., Kang, H. J., Stewart, H. J., & Amitay, S. (2015). Audiovisual integration in children listening to spectrally degraded speech. *Journal of Speech, Language, and Hearing Research*, *58*(1), 61–68.

Marian, V., Hayakawa, S., Lam, T., & Schroeder, S. (2018). Language experience changes audiovisual perception. *Brain Sciences*, *8*(5), 85.

Marques, L. M., Lapenta, O. M., Costa, T. L., & Boggio, P. S. (2016). Multisensory integration processes underlying speech perception as revealed by the McGurk illusion. *Language, Cognition, and Neuroscience*, *31*, 1115–1129.

Massaro, D. W. (1984). Children's perception of visual and auditory speech. *Child Development*, *5*, 1777–1788.

Massaro, D. W. (2017). The McGurk effect: Auditory visual speech perception's piltdown man. In Ouni, S., Davis, C., Jesse, A., & Beskow, J. (Eds.) *The 14th International Conference on Auditory-Visual Speech Processing. Stockholm, Sweden: KTH.*

Massaro, D. W., & Cohen, M. M. (2000). Tests of auditory–visual integration efficiency within the framework of the fuzzy logical model of perception. *The Journal of the Acoustical Society of America*, *108*(2), 784–789.

Massaro, D. W., Thompson, L. A., Barron, B., & Laren, E. (1986). Developmental changes in visual and auditory contributions to speech perception. *Journal of Experimental Child Psychology*, *41*, 93–113.

Mathôt, S., Schreij, D., & Theeuwes, J. (2012). Open sesame: An open-source, graphical experiment builder for the social sciences. *Behavior Research Methods*, *44*(2), 314–324.

McGettigan, C., Faulkner, A., Altarelli, I., Obleser, J., Baverstock, H., & Scott, S. K. (2012). Speech comprehension aided by multiple modalities: Behavioural and neural interactions. *Neuropsychologia*, *50*(5), 762–776.

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

Miller, J. L. (1994). On the internal structure of phonetic categories: a progress report. *Cognition*, *50*(1-3), 271–285.

Munhall, K. G., Gribble, P., Sacco, L., & Ward, M. (1996). Temporal constraints on the McGurk effect. *Perception & Psychophysics*, *58*, 351–362.

Nath, A. R., & Beauchamp, M. S. (2012). A neural basis for interindividual differences in the McGurk Effect, a multisensory speech illusion. *Neuroimage*, *59*, 781–787.

Navarra, J., Alsius, A., Soto-Faraco, S., & Spence, C. (2010). Assessing the role of attention in the audiovisual integration of speech. *Information Fusion*, *11*, 4–11.

Norrix, L. W., Plante, E., & Vance, R. (2006). Auditory-visual speech integration by adults with and without language-learning disabilities. *Journal of Communication Disorders*, *39*, 22–36.

Norrix, L. W., Plante, E., Vance, R., & Boliek, C. A. (2007). Auditory-visual integration for speech by children with and without specific language impairment. *Journal of Speech, Language, and Hearing Research*, *50*(6), 1639–1651.

Olasagasti, I., Bouton, S., & Giraud, A. L. (2015). Prediction across sensory modalities: A neurocomputational model of the McGurk effect. *Cortex*, *68*, 61–75.

Palmer, T. D., & Ramsey, A. K. (2012). The function of consciousness in multisensory integration. *Cognition*, *125*, 353–364.

Patterson, M. L., & Werker, J. F. (1999). Matching phonetic information in lips and voice is robust in 4.5-month-old infants. *Infant Behaviour & Development*, *22*, 237–247.

Patterson, M. L., & Werker, J. F. (2003). Two-month-old infants match phonetic information in lips and voice. *Developmental Science*, *6*, 191–196.

Pearl, D., Yodashkin-Porat, D., Katz, N., Valevski, A., Aizenberg, D., & Sigler, M. (2009). Differences in audiovisual integration, as measured by McGurk phenomenon, among adult and adolescent patients with schizophrenia and age-matched healthy control groups. *Comprehensive Psychiatry*, *50*(2), 186–192.

Pisoni, D. B., & Lazarus, J. H. (1974). Categorical and noncategorical modes of speech perception along the voicing continuum. *Journal of the Acoustical Society of America*, *55*(2), 328–333.

Pons, F., & Lewkowicz, D. (2014). Infant perception of audio-visual speech synchrony in familiar and unfamiliar fluent speech. *Acta Psychologica*, *149*, 142–147.

R Core Team (2019). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria.

Rosen, S. M. (1979). Range and frequency effects in consonant categorization. *Journal of Phonetics*, *7*(4), 393–402.

Rosenblum, L. D. (2008). Speech perception as a multimodal phenomenon. *Current Directions in Psychological Science*, *17*(6), 405–409.

Rosenblum, L. D., & Saldana, H. M. (1992). Discrimination tests of visually-influenced syllables. *Perception & Psychophysics*, *52*, 461–473.

Rosenblum, L. D., & Saldana, H. M. (1996). An audiovisual test of kinematic primitives for visual speech perception. *Journal of Experimental Psychology: Human Perception and Performance*, *22*, 318–331.

Rosenblum, L. D., Schmuckler, M. A., & Johnson, J. A. (1997). The McGurk effect in infants. *Perception & Psychophysics*, *59*(3), 347–357.

Ross, L. A., Molholm, S., Blanco, D., Gomez Ramirez, M., Saint Amour, D., & Foxe, J. J. (2011). The development of multisensory speech perception continues into the late childhood years. *European Journal of Neuroscience*, *33*, 2329–2337.

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

Sams, M., Manninen, P., Surakka, V., Helin, P., & Katto, R. (1998). McGurk effect in Finnish syllables, isolated words, and words in sentences: Effects of word meaning and sentence context. *Speech Communication*, *26*, 75–87.

Sánchez-García, C., Alsius, A., Enns, J. T., & Soto-Faraco, S. (2011). Cross-modal prediction in speech perception. *PloS one*, *6*(10), e25198.

Schorr, E. A., Fox, N. A., van Wassenhove, V., & Knudsen, E. I. (2005). Auditory-visual fusion in speech perception in children with cochlear implants. *Proceedings of the National Academy of Sciences*, *102*, 18748–18750.

Schouten, B., Gerrits, E., & van Hessen, A. (2003). The end of categorical perception as we know it. *Speech Communication*, *41*, 71–80.

Schwartz, J. L. (2010). A reanalysis of McGurk data suggests that audiovisual fusion in speech perception is subject-dependent. *Journal of the Acoustical Society of America*, *127*(3), 1584–1594.

Sekiyama, K. (1997). Cultural and linguistic factors in audiovisual speech processing: The McGurk effect in Chinese subjects. *Perception & Psychophysics*, *59*(1), 73–80.

Sekiyama, K., & Burnham, D. (2004). Issues in the development of auditory-visual speech perception: adults, infants, and children. In *Interspeech, p. 1137–1140*.

Sekiyama, K. (2008). Burnham, d. *Impact of language on development of auditory-visual speech perception. Developmental Science*, *11*, 306–320.

Sekiyama, K., Soshi, T., & Sakamoto, S. (2014). Enhanced audiovisual integration with aging in speech perception: a heightened McGurk effect in older adults. *Frontiers in Psychology*, *5*, 323.

Soto-Faraco, S., & Alsius, A. (2009). Deconstructing the McGurk-MacDonald illusion. *Journal of Experimental Psychology: Human Perception and Performance*, *35*(2), 580–587.

Strand, J., Cooperman, A., Rowe, J., & Simenstad, A. (2014). Individual differences in susceptibility to the McGurk effect: Links with lipreading and detecting audiovisual incongruity. *Journal of Speech, Language, and Hearing Research*, *57*(6), 2322–2331.

Sumby, W. H., & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *The Journal of the Acoustical Society of America*, *26*, 212–215.

Summerfield, Q. (1979). Use of visual information for phonetic perception. *Phonetica*, *36*(4-5), 314–331.

Tomalski, P. (2015). Developmental trajectory of audiovisual speech integration in early infancy: A review of studies using the McGurk paradigm. *Psychology of Language and Communication*, *19*(2), 77–100.

Toscano, J. C., & Lansing, C. R. (2019). Age-related changes in temporal and spectral cue weights in speech. *Language and Speech*, *62*, 61–79.

Toscano, J. C., McMurray, B., Dennhardt, J., & Luck, S. J. (2010). Continuous perception and graded categorization electrophysiological evidence for a linear relationship between the acoustic signal and perceptual encoding of speech. *Psychological Science*, *21*(10), 1532–1540.

Tremblay, C., Champoux, F., Voss, P., Bacon, B., Lepore, F., & Theoret, H. (2007). Speech and non-speech audio-visual illusions: A developmental study. *PLoSOne*, *2*(1), e742.

Tuomainen, J., Andersen, T. S., Tiippana, K., & Sams, M. (2005). Audio-visual speech perception is special. *Cognition*, *96*, B13–B22.

Van Engen, K. J., Xie, Z., & Chandrasekaran, B. (2017). Audiovisual sentence recognition not predicted by susceptibility to the McGurk effect. *Attention, Perception, & Psychophysics*, *79*(2), 396–403.

Viswanathan, N., Magnuson, J. S., & Fowler, C. A. (2010). Compensation for coarticulation: Disentangling auditory and gestural theories of perception of coarticulatory effects in speech. *Journal of Experimental Psychology: Human Perception and Performance*, *36*(4), 1005–1015.

Vogel, E. K., & Awh, E. (2008). How to exploit diversity for scientific gain: Using individual differences to constrain cognitive theory. *Current Directions in Psychological Science*, *17*(2), 171–176.

Wightman, F., Kistler, D., & Brungart, D. (2006). Informational masking of speech in children: Auditory-visual integration. *The Journal of the Acoustical Society of America*, *119*, 3940–3949.

Wilson, A., Alsius, A., Pare, M., & Munhall, K. (2016). Spatial frequency requirements and gaze strategy in visual-only and audiovisual speech perception. *Journal of Speech, Language, and Hearing Research*, *59*, 601–615.

Windmann, S. (2004). Effects of sentence context and expectation on the McGurk illusion. *Journal of Memory and Language*, *50*(2), 212–230.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.