

Variability and stability in the McGurk effect: contributions of participants, stimuli, time, and response type

Debshila Basu Mallick · John F. Magnotti ·
Michael S. Beauchamp

© Psychonomic Society, Inc. 2015

Abstract In the McGurk effect, pairing incongruent auditory and visual syllables produces a percept different from the component syllables. Although it is a popular assay of audiovisual speech integration, little is known about the distribution of responses to the McGurk effect in the population. In our first experiment, we measured McGurk perception using 12 different McGurk stimuli in a sample of 165 English-speaking adults, 40 of whom were retested following a one-year interval. We observed dramatic differences both in how frequently different individuals perceived the illusion (from 0 % to 100 %) and in how frequently the illusion was perceived across different stimuli (17 % to 58 %). For individual stimuli, the distributions of response frequencies deviated strongly from normality, with 77 % of participants almost never or almost always perceiving the effect ($\leq 10\%$ or $\geq 90\%$). This deviation suggests that the mean response frequency, the most commonly reported measure of the McGurk effect, is a poor measure of individual participants' responses, and that the assumptions made by parametric statistical tests are invalid. Despite the substantial variability across individuals and stimuli, there was little change in the frequency of the effect between initial testing and a one-year retest (mean change in frequency = 2 %; test-retest correlation, $r = 0.91$). In a second experiment, we replicated our findings of high variability using eight new McGurk stimuli and tested the effects of open-choice versus forced-choice responding. Forced-choice responding resulted in an estimated 18 % greater frequency of the McGurk effect but similar levels of interindividual variability. Our results highlight the importance of examining

individual differences in McGurk perception instead of relying on summary statistics averaged across a population. However, individual variability in the McGurk effect does not preclude its use as a stable measure of audiovisual integration.

Keywords Individual differences · Audiovisual integration · McGurk effect · Speech perception

Humans use information from both the auditory modality (the sound of the talker's voice) and the visual modality (the sight of the talker's face) to understand spoken language. The *McGurk effect* is an illusion that demonstrates the importance of the visual modality for speech perception: Pairing an auditory syllable with an incongruent visual syllable produces the percept of a third syllable, different from both the auditory and visual syllables (McGurk & MacDonald, 1976). Because of its simplicity, the McGurk effect has been used as a measure of audiovisual integration in healthy children (Nath, Fava, & Beauchamp, 2011; Tremblay, Champoux, Bacon, Lepore & Théoret, 2007a); in children and adults (Erdener, Sekiyama & Burnham, 2010; Tremblay, Champoux, Voss, Bacon, Lepore & Théoret, 2007b); in clinical groups, such as individuals with autism spectrum disorder (Irwin, Tornatore, Brancazio, & Whalen, 2011; Woynaroski et al., 2013); and to examine the neural substrates of speech perception (Beauchamp, Nath, & Pasalar, 2010; Keil, Müller, Ihssen, & Weisz, 2012; McKenna Benoit, Raji, Lin, Jääskeläinen, & Stufflebeam, 2010; Nath & Beauchamp, 2012; Skipper, van Wassenhove, Nusbaum, & Small, 2007).

Fundamental questions about the McGurk effect remain unanswered. Most importantly, the distribution of responses to McGurk stimuli in the population and their stability over time is poorly understood. In the initial description of the effect (McGurk & MacDonald, 1976), 98 % of adult participants reported an illusory "da" percept when an auditory "ba"

D. Basu Mallick
Department of Psychology, Rice University, Houston, TX, USA
e-mail: debshila@rice.edu

J. F. Magnotti · M. S. Beauchamp (✉)
Department of Neurosurgery, Baylor College of Medicine,
1 Baylor Plaza, Houston 77030, TX, USA
e-mail: Michael.Beauchamp@bcm.edu

was paired with a visual “ga,” whereas a follow-up study reported a frequency of only 64 % for the same combination (MacDonald & McGurk, 1978). At least three obvious possibilities may explain such differing estimates of the frequency of the McGurk effect. First, because the original McGurk stimuli are no longer available, different studies use different McGurk stimuli, often created in the laboratory solely for that study. Second, there are substantial individual differences in the frequency of the McGurk effect, from 0 % to 100 % across different participants (Keil et al., 2012; McKenna Benoit et al., 2010; Nath & Beauchamp, 2012; Sekiyama, Braida, Nishino, Hayashi, & Tuyo, 1995; Stevenson, Zemtsov, & Wallace, 2012; Tremblay et al., 2007a). Third, different studies use different experimental procedures, and a procedure that incorporates experimenter expectations (“did the stimulus sound like da?”) might give different results than one that does not (“what did the stimulus sound like?”) (Colin, Radeau, & Deltenre, 2005; Orne, 1962). In order to assess the possible contributions of differences in the stimuli, participants, and procedures to the variation in the published estimates of McGurk frequency, we tested 360 individuals, 20 different McGurk stimuli, and open-choice and forced-choice experimental procedures. For the McGurk effect to be a useful measure of audiovisual integration, it must not vary greatly within individuals from

day to day. To assess the stability of the effect within individuals, we tested 40 individuals across a one-year interval.

Method

Experiment 1

Undergraduate students from Rice University participated in the study for psychology course credit ($N = 165$: 106 female, 59 male; mean age = 19 years). The participants gave written informed consent under an experimental protocol approved by the Rice University Institutional Review Board. After one year, participants were invited to return for an additional testing session; 40 participants (29 female, 11 male) did so. All of the participants had normal or corrected-to-normal vision and no hearing impairments.

Stimuli The stimuli consisted of 12 McGurk syllable pairs that had been used in previously published studies or were freely available on the Internet (Table 1). Seven of the McGurk stimuli were tested with all $N = 165$ participants, one stimulus was tested with a subset of $N = 88$ participants, and four stimuli were tested with a subset of $N = 66$. For the one-year retest session, only stimuli used at both test and retest were

with the auditory (A) and visual (V) components comprising each stimulus, and the number of participants tested with each stimulus in Experiment 1 (N)

ID	A + V	N	Source
1	ba + ga	165	Nath & Beauchamp (2012); Openwetware ^a
2	ba + ga	66	Sekiyama et al. (1995)
3	baba + gaga	165	http://youtu.be/tUff0672xAOU
4	baba + gaga	165	http://youtu.be/5Lq26mgFpOc
5	ba + ga	66	Kaoru Sekiyama, Kumamoto University
6	ba + ga	88	Quinto, Thompson, Russo, & Trehab (2010)
7	ba + ga	66	Kaoru Sekiyama, Kumamoto University
8	ba + ga	66	Sekiyama et al. (1995)
9	baba + gaga	165	http://youtu.be/aFPtc8BVdJk
10	ba + ga	165	http://youtu.be/rIWrnJH2jAY
11	ba + ga	165	Doğu Erdener, Middle East Technical University
12	ba + ga	165	http://youtu.be/jtsfidRq2tw
C1	ba + ba	165	Nath & Beauchamp (2012); Openwetware
C2	ga + ga	165	Nath & Beauchamp (2012); Openwetware
C3	da + da	165	Nath & Beauchamp (2012); Openwetware
A1	ba	165	Openwetware
A2	baba	165	Openwetware
A3	ba	165	Openwetware
A4	baba	165	Openwetware

^a The stimuli on Openwetware are available from <http://openwetware.org/wiki/Beauchamp:McGurkStimuli>.

compared. All stimuli were edited to have a duration of 1.5–2.0 s and were sized at 640 × 480 pixels. The control stimuli consisted of eight auditory-only syllables and eight audiovisual (non-McGurk) syllables. All stimuli were presented using the Psychophysics Toolbox (Brainard, 1997; Pelli, 1997).

Experimental procedure Participants sat at a desk in a quiet testing room about 50 cm away from a 15-in. laptop screen (1, 440 × 900 resolution) and wore noise-attenuating headphones. First, participants were presented with auditory-only control stimuli after receiving the following instructions: “You will hear recordings of different people talking. After each recording, repeat loudly and clearly what the person said. If you are not sure, take your best guess. There are no right or wrong answers.” Eight auditory-only stimuli were played five times in random order, and responses were entered into a spreadsheet by the research assistants.

Second, participants were presented with audiovisual stimuli after receiving the following instructions: “You will see videos of different people talking. Please watch the screen. After each video, wait until the gray screen appears. As soon as it appears, repeat loudly and clearly what the person said. If you are not sure, take your best guess. There are no right or wrong answers.” The instructions were designed to be modality neutral, so as not to bias participants toward auditory or visual responses. Audiovisual stimuli were presented ten times each in random order.

Data analysis Responses to the McGurk stimuli were divided into four mutually exclusive categories (McGurk & MacDonald, 1976): fusion responses (“da” or “tha”), auditory responses (“ba”), visual responses (“ga”), and other (e.g., “va”). Stimuli 3, 4, and 9 consisted of two syllables (auditory “baba” paired with visual “gaga”). For these stimuli, a half-point was assigned to each syllable: For instance, the response “dada” was scored as 1.0 as a fusion response, whereas “daba” was scored as 0.5 fusion 0.5 auditory. All data were analyzed using R statistical software (R Development Core Team, 2014).

Experiment 2

Stimuli Because the stimuli in Experiment 1 were selected from previous studies, they had been created in many different laboratories and varied along a number of dimensions, including auditory and visual quality and the size of the face within the video frame. To minimize the effect of these potential confounds, we created an additional stimulus set consisting of eight McGurk stimuli (labeled 2.1–2.8) recorded from four male and four female talkers; a different talker was used to record three audiovisual control stimuli (congruent “ba,” “da,” and “ga”). The stimuli in Experiment 2 are available for download at <http://openwetware.org/wiki/Beauchamp:McGurkStimuli>.

Experimental procedure and participants Data were collected using an online data collection service (Amazon Mechanical Turk). The eight McGurk stimuli were each presented ten times, and the control stimuli were presented five times, all randomly interleaved. One group of participants ($N = 110$: 38 female, 72 male) used an open-choice response format, typing their percept following the presentation of each stimulus; these responses were coded as in Experiment 1. A second group of participants ($N = 117$: 55 female, 62 male) made a forced choice, selecting among three possibilities corresponding to the auditory (“ba”), visual (“ga”), or McGurk fusion percept (“da” or “tha”) with a mouse click. Six participants who did not report their gender were used for all analyses except gender comparisons, six participants were excluded for failing to complete the task as instructed, and 32 participants completed both experiments (total unique $N = 195$).

Experiment 3

To determine the accuracy of unisensory syllable perception, we extracted the auditory and visual components from the eight McGurk stimuli used in Experiment 2 and tested them using the same forced-choice procedure as in Experiment 2. Participants ($N = 50$: 21 female, 29 male; seven had also participated in Experiment 2) were presented with each of auditory-only and visual-only “ba,” “da,” and “ga” from eight talkers (48 unique stimuli). Each stimulus was repeated twice (96 trials) and randomly interleaved.

Results

Experiment 1, Session 1

Across 165 participants, there was a high degree of variability in the frequency of the McGurk percepts, ranging from 0 % (no illusory percepts reported for any presentation of any stimulus) to 100 % (illusory percept reported on every presentation of every stimulus), covering many values in between (Fig. 1a).

Within participants, we did not observe consistent response frequencies across stimuli (Fig. 1b). Although Participant 108 had a mean frequency of 50 %, only one stimulus had a frequency near this value (Fig. 1b). The mean of 50 % resulted from six stimuli with low frequencies ($\leq 20\%$) and five stimuli with high frequencies ($\geq 90\%$).

To determine whether this all-or-none McGurk perception was typical, we examined the response distributions for individual stimuli. As with participants, we found a high degree of variability in McGurk frequencies for different stimuli, with the most effective stimulus eliciting over 3 times as many McGurk responses as the least effective one (range from 17 % for Stimulus 1 to 58 % for Stimulus 12; Fig. 2a).

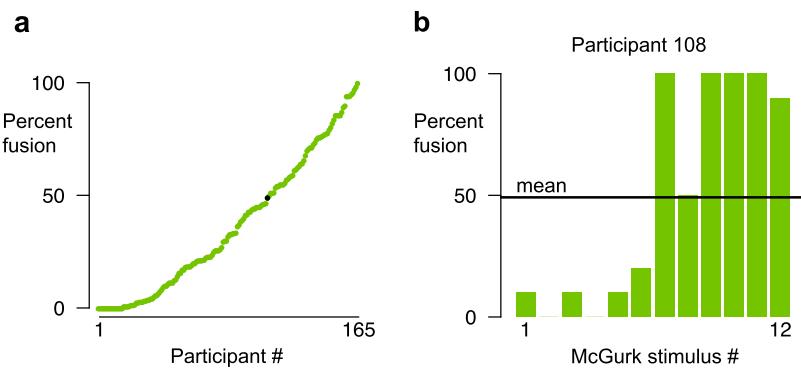


Fig. 1 Responses to McGurk stimuli during Session 1 of Experiment 1. (a) Mean frequencies of McGurk responses for 165 participants (one symbol per participant, ordered by increasing frequency; data from Exp. 1, Session 1). Fusion responses ranged from 0 % to 100 %. The black

symbol shows Participant 108. (b) Frequencies of McGurk responses of Participant 108 to each of the 12 McGurk stimuli. The black line shows the mean frequency across stimuli for this participant; most stimuli are far above or below the mean.

Within a given stimulus, the response frequencies of individual participants tended to be all-or-none. For Stimulus 9

(Fig. 2b), 49 % of the participants never or almost never reported the illusion (frequency $\leq 10 \%$), and 39 % of the

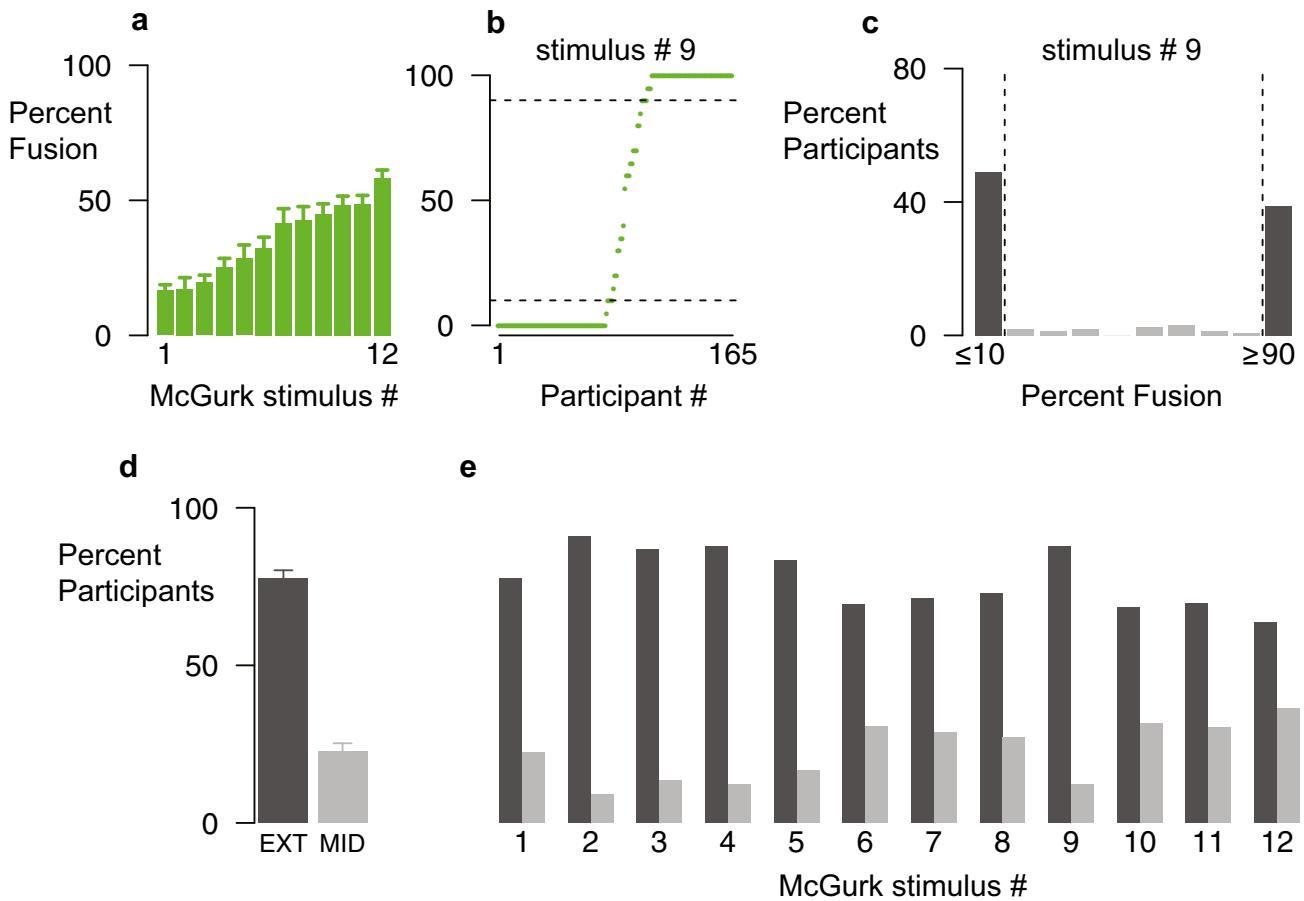


Fig. 2 Distributions of McGurk response percentages across stimuli. (a) Mean percentages of McGurk responses for 12 different stimuli (error bars show SEMs; data from Exp. 1, Session 1). (b) Percentage of McGurk responses for Stimulus # 9 in each of the 165 participants (one symbol per participant, ordered by increasing percentages). Dashed horizontal lines show the extremes of the distribution ($\leq 10 \%$, $\geq 90 \%$). (c) Percentages of McGurk responses for Stimulus 9, plotted as the percentage of

participants within each 10 % frequency bin. Dark bars highlight participants in the extremes of the distribution. (d) Average percentages of participants across stimuli in the extremes of the distribution (EXT, dark bar) and the middle ($>10 \%$ and $<90 \%$) of the distribution (MID, light bar). Error bars indicate SEMs. (e) Percentages of participants in the extremes and in the middle of the distribution for each individual stimulus.

participants always or almost always reported the illusion (frequency $\geq 90\%$). Plotting the percentage of participants with each McGurk frequency for this stimulus illustrates the concentration of participants at the extreme values (Fig. 2c). Across stimuli, 77 % of participants almost never or almost always perceived the McGurk effect ($\leq 10\%$ or $\geq 90\%$), whereas only 23 % had intermediate frequencies ($> 10\%$ but $< 90\%$; Fig. 2d). This pattern held for each individual stimulus (Fig. 2e) and is the opposite of the pattern expected in a normal distribution, in which most samples are concentrated near the mean. A Shapiro–Wilk (1965) test of normality rejected normality for each stimulus distribution (all $p < 10^{-8}$). Distributions that are non-normal may still be unimodal, but a dip test (Hartigan & Hartigan, 1985) rejected the hypothesis that any of the distributions were unimodal (all $p < 10^{-16}$), meaning that other measures of central tendency, such as the mode and median, are also poor descriptors. For instance, for Participant 108, the mean frequency was 50 %, the modal frequency was 100 %, and the median frequency was 35 %, none of which describe the participant's pattern of responses well (Fig. 1b).

No gender difference emerged in the frequency of the McGurk effect (mean across genders, 36.8 %; females: 36.7 %, $SEM = 2.9\%$; males: 37.2 %, $SEM = 4.1\%$; Kolmogorov–Smirnov $D = 0.10, p = 0.83$). When participants did not report the fusion percept (“da” or “tha”), they reported the auditory component of the stimulus (“ba,” 51 %), the visual component of the stimulus (“ga,” 1 %), or some other percept (10 %). The other percepts included “ah” (3 %), “fa” (3 %), “ta” (2 %), “pa” (1 %), “la” (0.6 %), “va” (0.5 %), and 13 other percepts (each reported less than 0.5 % of the time). Participants were at ceiling accuracy for control stimuli consisting of auditory-only syllables (mean = 98 %) and congruent audiovisual syllables (mean = 99 %).

Experiment 1, Session 2

A total of 40 participants returned for a second testing session at least one year after the first testing session. We found no change in the frequency of the McGurk effect within individuals between the two sessions [mean difference in each individual = 2 %; paired t -test, $t(39) = 1.54, p = 0.13$], resulting in a high correlation between test and retest (Fig. 3; $r = 0.91, p = 10^{-16}$; Cronbach's $\alpha = 0.96$ [Cronbach, 1951]). There was no gender difference in the frequency of McGurk responses at retest (females: mean = 43.3 %, $SEM = 6.0\%$; males: mean = 47.4 %, $SEM = 10.1\%$; $D = 0.19, p = 0.95$).

Experiment 2: Effect of choice type and replication

In Experiment 2, we presented a set of eight new McGurk stimuli to two different groups of participants. The first group used the same open-choice design as in Experiment 1, but the

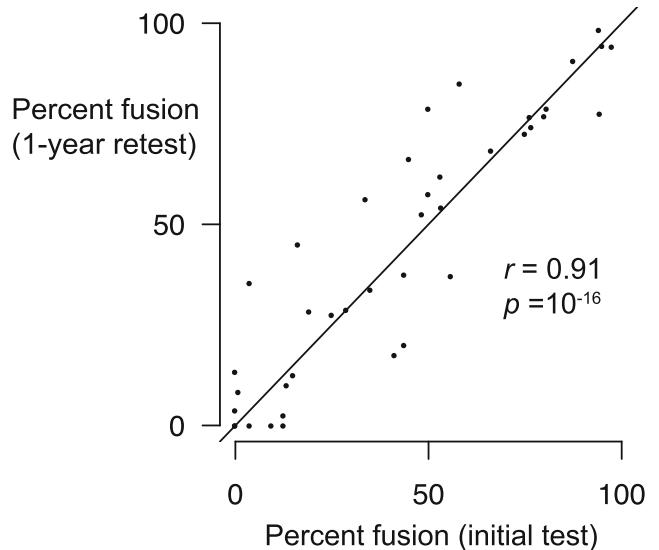


Fig. 3 Stability of McGurk responses after a one-year interval: Mean percentages of McGurk responses for each participant for the initial test (Exp. 1, Session 1) and at retest (Exp. 1, Session 2) one year later. Each point represents one participant.

second group made a three-alternative forced choice (corresponding to the auditory component of the stimulus, the visual component of the stimulus, or the illusory McGurk percept). The forced-choice group was much more likely to report the McGurk effect than the open-choice group (69 % vs. 42 %, Kolmogorov–Smirnov $D = 0.36, p = 10^{-7}$; see Fig. 4a). This increase was consistent across all eight stimuli (Fig. 4b).

Replicating the results from Experiment 1, we found high variability across participants (range from 0 % to 100 %) in both the open-choice (Fig. 4c) and forced-choice (Fig. 4d) groups. Replicating Experiment 1, most participants were found in the extremes of the distribution, both averaged across stimuli (Fig. 4e) and for each individual stimulus (Fig. 4f; data combined across choice groups). A Shapiro–Wilk (1965) test of normality rejected normality for each stimulus distribution (all $p < 10^{-13}$), and a dip test (Hartigan, & Hartigan, 1985) rejected the hypothesis that any of the distributions were unimodal (all $p < 10^{-16}$).

Replicating the results from Experiment 1, we also found high variability across stimuli, ranging from 30 % to 52 % for open choice and 57 % to 80 % for forced choice. Many possible factors could contribute to the differences in McGurk frequencies that we observed across stimuli in Experiment 1, including the size of the face within the video frame, the auditory sound quality, the use of single versus double syllables, and talker gender. The McGurk stimulus set for Experiment 2 was created with four female and four male talkers, allowing us to measure the effects of talker gender and its interaction with the participant and task factors, while holding other stimulus factors constant. We fit a linear mixed-effects model to the behavioral data, with choice type (open vs. forced), talker gender, participant gender, and their interactions as fixed

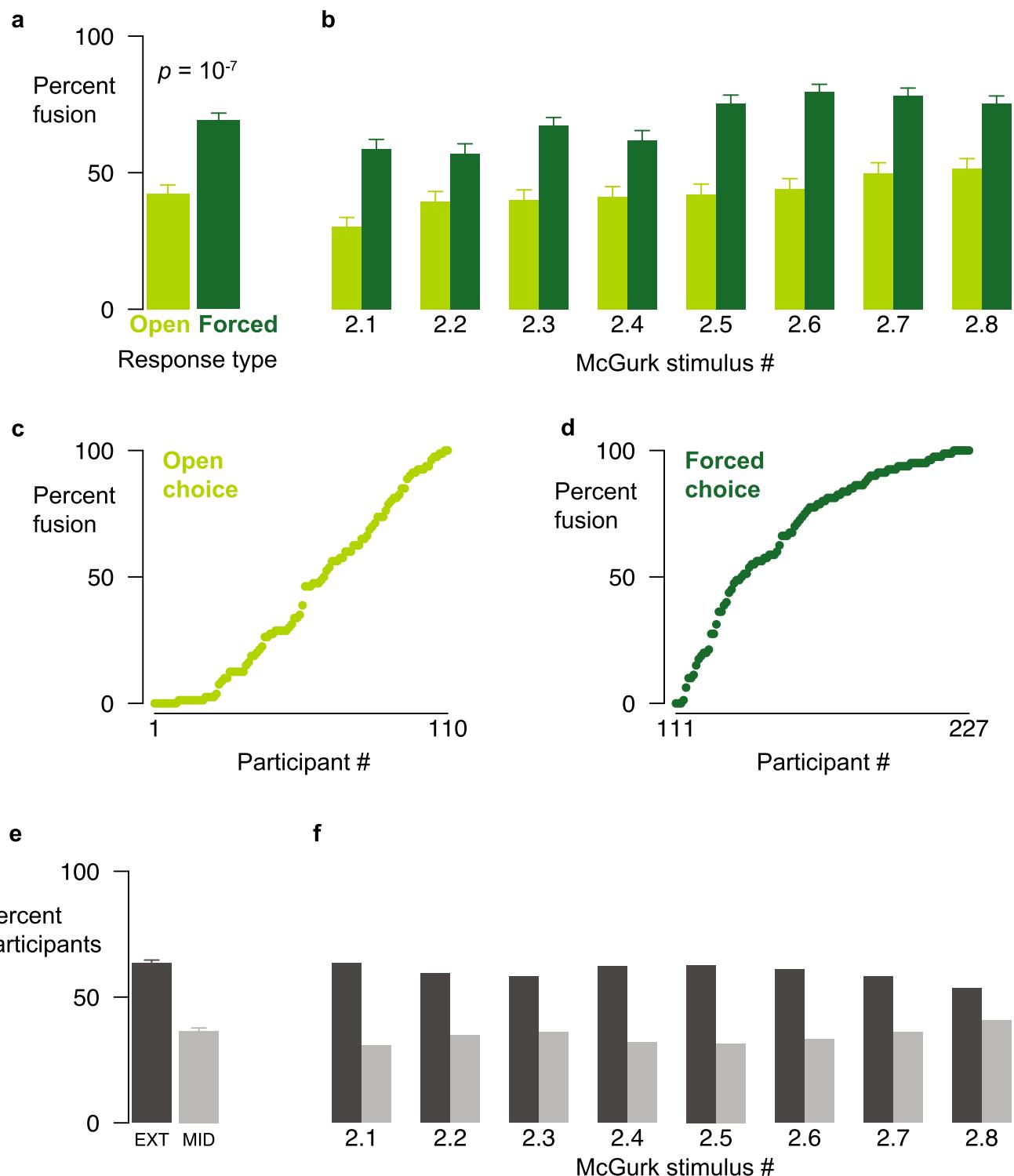


Fig. 4 Data from Experiment 2, comparing open-choice and forced-choice responding. (a) Percentages of fusion responses averaged across stimuli for open-choice (Open) and three-alternative forced-choice (Forced) responding, tested in two different groups of participants. Error bars indicate SEMs. (b) Percentages of fusion responses for each individual stimulus. (c) Percentage of fusion responses for each individual

participant in the open-choice group. (d) Percentage of fusion responses for each individual participant in the forced-choice group. (e) Average percentages of participants across stimuli at the extremes (EXT, dark bar) and in the middle (>10 % and <90 %; MID, light bar) of the distribution. Error bars show SEMs. (f) Percentages of participants in the extremes and in the middle of the distribution for each individual stimulus.

effects, and participant and stimulus as random effects. Using Satterthwaite approximations to test the significance of the

model coefficients, the only large effect was the main effect of choice type [estimated 18.0 % higher for forced choice, SE

$= 3.2\%;$ $t(1754) = 5.6, p = 10^{-8}$]. There was no main effect of participant gender [8.1 % lower for males, $SE = 5.1\%$; $t(292.3) = -1.6, p = .12$] or talker gender [2.6 % higher for male talkers, $SE = 6.0\%$; $t(7.8) = 0.43, p = 0.68$], and only weak interactions between participant gender and choice type [5.8 % lower for male participants in forced choice, $SE = 4.1\%$; $t(1753) = -1.4, p = .15$], talker gender and choice type [6.8 % lower for male talkers in forced choice, $SE = 3.0\%$; $t(1562) = -2.3, p = 0.02$], and talker gender and participant gender [3.4 % lower for male participants viewing male talkers, $SE = 2.9\%$; $t(1562) = -1.2, p = 0.24$]. The three-way interaction was also weak [3.2 % lower for male participants viewing male talkers in forced choice, $SE = 3.9\%$; $t(1562) = -0.08, p = 0.94$]. Eliminating participant gender and stimulus gender from the model did not significantly change its predictive accuracy [full model, root mean squared error (RMSE) = 19.1 %; reduced model, RMSE = 19.3 %; mean difference = 0.05 %; paired t -test: $t(1767) = -1.14, p = .25$].

Experiment 3

Because the stimuli in Experiment 1 were selected from previous studies, we did not have access to the talkers in each stimulus speaking other syllables. Therefore, while creating the McGurk stimuli for Experiment 2, we recorded the same eight talkers speaking “ba,” “ga,” and “da.” We presented the auditory and visual components of these stimuli in Experiment 3 using the three-alternative forced-choice design used in Experiment 2. Identification of the auditory-only syllables was at ceiling (mean accuracy = 97 %, $SD = 4\%$), whereas identification of the visual-only syllables was significantly worse (80 %, $SD = 10\%$) [paired t -test: $t(49) = 13, p < 10^{-16}$]. Accuracy for the “ga” visual-only syllable was especially low (58 %, as compared with 96 % for “ba” and 86 % for “da”), with high variability across talkers (range from 5 % to 77 %) and participants (range from 6 % to 88 %).

Discussion

Across 360 individuals and 20 different McGurk stimuli, we observed an astonishing diversity of responses to the illusion. In Experiment 1, we found that some participants never perceived the illusion across more than a hundred presentations (frequency of 0 %), whereas others perceived it every single time (frequency of 100 %). Similarly large variability was observed for stimuli created from different talkers (range of 17 % to 58 %). Experiment 2 replicated these findings and showed that manipulating response type also significantly alters the frequency of the McGurk effect, with forced-choice responding increasing the frequency of McGurk perception by an estimated 18 %, as compared with open choice for identical stimuli, similar to the findings of Colin et al. (2005). Together, these

results demonstrate that differences in participants, stimuli, and experimental paradigms all contribute to the wide range of published estimates of the frequency of the McGurk effect. The high variability in the effect suggests that caution is necessary when comparing McGurk frequencies across groups or across studies in which any of these factors vary.

Although the McGurk frequencies for individual participants were evenly distributed across the range from 0 % to 100 %, within each stimulus we found an all-or-none pattern of responding. For the individual stimuli in Experiment 1, 77 % of participants almost never or almost always perceived the McGurk effect ($\leq 10\%$ or $\geq 90\%$). Therefore, using the mean and standard deviation to characterize the frequency of the McGurk effect can lead to errors in inference, both conceptual and statistical (Gravetter & Wallnau, 2006). For example, for Stimulus 9, the mean frequency of McGurk responses was 45 %, but only 7 % of participants had a McGurk frequency near the mean (frequencies of 25 % to 65 %).

The non-normality of the distribution of McGurk responses violates the assumptions of parametric statistics such as t tests and reduces their ability to detect differences between groups. For instance, Stimulus 10 had many more participants in the middle of the distribution than did Stimulus 9 (32 % vs. 12 %; Fig. 2e). Although a t test showed no difference between the stimuli (mean frequencies of $45\% \pm 47\%$ vs. $48\% \pm 42\%$) [$t(164) = -1.31, p = 0.2$], the Kolmogorov–Smirnov test, which does not assume any distribution, successfully detected the difference ($D = 0.22, p = 0.0004$); similar problems could prevent the t test from detecting mean differences between two groups. This example shows that researchers employing the McGurk effect should be wary of using mean frequencies to examine group differences.

A more fundamental problem raised by our results is that raw measures of McGurk frequency confound individual differences and stimulus differences. To extract the individual differences (which are typically of greater interest), one solution would be to apply the noisy encoding of disparity (NED) model of the McGurk effect (Magnotti & Beauchamp, 2014). The NED model separately estimates individual parameters (disparity threshold and sensory noise) that can be used to make participant and group comparisons that are unaffected by stimulus differences. The NED model also provides an explanation for the non-normality of the McGurk response data. Stimuli are characterized by a disparity between the auditory and visual speech cues. If the perceived stimulus disparity falls below an individual’s disparity threshold, the individual infers that the auditory and visual speech cues arise from the same talker and integrate them, resulting in the McGurk percept, whereas if a stimulus falls above the threshold, the individual infers that the cues arise from different talkers and does not integrate them, resulting in a percept of the auditory component of the stimulus. As perceived stimulus

disparity decreases below an individual's threshold, there is a rapid transition from never perceiving the illusion to always perceiving it. Only in the rare case that the stimulus disparity almost exactly matches the individual's threshold will the illusion be perceived on some presentations but not others. The NED model does not specify the source of the disparity between the auditory and visual speech cues. Participants show remarkably high tolerances for temporal asynchrony between the auditory and visual components of speech (Magnotti, Ma, & Beauchamp, 2013; Munhall, Gribble, Sacco, & Ward, 1996), and perceive the McGurk effect even if the voice and the face are of different genders (Green, Kuhl, Meltzoff, & Stevens, 1991). A comprehensive study of many acoustic and visual properties of McGurk stimuli showed that they accounted for about half of the variability in the frequency of the effect across stimuli and participants (Jiang & Bernstein, 2011).

Previous studies have characterized individuals as being strong or weak perceivers of the McGurk effect in order to show group differences in mental imagery (Berger & Ehrsson, 2013) or brain activity in adults (Nath & Beauchamp, 2012) and children (Nath et al., 2011). Our data suggest that this is reasonable for any particular stimulus, since the responses to all stimuli were characterized by distributions in which most participants were found in the extremes (either almost always or almost never perceiving the illusion). However, the classification into strong and weak perceivers is entirely dependent on the stimulus chosen and the behavioral paradigm used. The weakest stimulus from Experiment 1 would classify 12 % of individuals as strong perceivers (using a classification threshold of 50 %), whereas the strongest stimulus from Experiment 2 using a forced-choice response would classify 84 % of individuals as strong perceivers (using the same classification threshold).

Our results confirm and extend a number of results from previous studies that have used smaller sample sizes. We did not find an effect of talker gender or of participant gender, consistent with previous reports for McGurk syllables (Irwin, Whalen, & Fowler, 2006) and visual-only phonemes (Strelnikov et al., 2009). Our visual-only results confirm that visual "da" and visual "ga" are easily confusable, whereas visual "ba" is distinct (Binnie, Montgomery & Jackson, 1974; Erber, 1975; Fisher, 1968; Lucey, Martin, & Sriradharan, 2004). Although we found a large effect of response type, we did not examine the effect of the task instructions. Our instructions were designed to be modality neutral, so as not to bias participants toward any particular response. It is possible that different task instructions could change the frequency of the McGurk effect, adding an additional source of variability across studies.

Although we found high variability in our examination of the McGurk effect, individual participants' McGurk frequencies were stable over a one-year period. We observed an $r = 0.91$ in 40 individuals over a 1 year interval, similar to the

findings of Strand, Cooperman, Rowe and Simenstad (2014) ($r = 0.77$ over a 2-month test-retest window in 58 individuals). This provides reassurance for longitudinal examinations of the McGurk effect and studies that correlate the frequency of the McGurk effect with brain activity (e.g., McKenna Benoit et al., 2010; Nath & Beauchamp, 2012; Nath et al., 2011), clinical status (e.g., Hamilton, Shenton, & Coslett, 2006; Pearl et al., 2009; Stevenson et al., 2014; Woynaroski et al., 2013), or other behavioral measures (e.g., Berger & Ehrsson, 2013; Stevenson et al., 2012; Tremblay et al., 2007a).

Author note This research was supported by NIH R01NS065395. Haley Lindsay, Cara Miekka, Joanne Guidry, and Alexandra Hernandez assisted with data collection. The authors report no conflicts of interest.

References

- Beauchamp, M. S., Nath, A. R., & Pasalar, S. (2010). fMRI-guided transcranial magnetic stimulation reveals that the superior temporal sulcus is a cortical locus of the McGurk effect. *Journal of Neuroscience*, 30, 2414–2417.
- Berger, C. C., & Ehrsson, H. H. (2013). Mental imagery changes multi-sensory perception. *Current Biology*, 23, 1367–1372.
- Binnie, C. A., Montgomery, A. A., & Jackson, P. L. (1974). Auditory and visual contributions to the perception of consonants. *Journal of speech, language, and hearing research*, 17, 619–630.
- Brainard, D. H. (1997). The Psychophysics Toolbox. *Spatial Vision*, 10, 433–436. doi:10.1163/156856897X00357
- Colin, C., Radeau, M., & Deltenre, P. (2005). Top-down and bottom-up modulation of audiovisual integration in speech. *European Journal of Cognitive Psychology*, 17(4), 541–560.
- Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16, 297–334. doi:10.1007/BF02310555
- Erber, N. P. (1975). Auditory-visual perception of speech. *Journal of Speech and Hearing Disorders*, 40(4), 481–492.
- Erdener, D., Sekiyama, K., & Burnham, D. (2010, August). *The development of auditory-visual speech perception across languages and age*. Paper presented at the 20th International Congress on Acoustics, Sydney, New South Wales, Australia.
- Fisher, C. G. (1968). Confusions among visually perceived consonants. *Journal of Speech, Language, and Hearing Research*, 11(4), 796–804.
- Gravetter, F., & Wallnau, L. (2006). *Statistics for the behavioral sciences*. Belmont, CA: Thomson Wadsworth.
- 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 & Psychophysics*, 50, 524–536.
- Hamilton, R. H., Shenton, J. T., & Coslett, H. B. (2006). An acquired deficit of audiovisual speech processing. *Brain and Language*, 98, 66–73.
- Hartigan, J., & Hartigan, P. (1985). The dip test of unimodality. *Annals of Statistics*, 13, 70–84.
- Irwin, J. R., Tornatore, L. A., Brancazio, L., & Whalen, D. H. (2011). Can children with autism spectrum disorders "hear" a speaking face? *Child Development*, 82, 1397–1403.
- Irwin, J. R., Whalen, D. H., & Fowler, C. A. (2006). A sex difference in visual influence on heard speech. *Perception & Psychophysics*, 68, 582–592. doi:10.3758/BF03208760

- 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, 1193–1209. doi:10.1037/a0023100
- Keil, J., Müller, N., Ihssen, N., & Weisz, N. (2012). On the variability of the McGurk effect: Audiovisual integration depends on prestimulus brain states. *Cerebral Cortex*, 22, 221–231.
- Lucey, P., Martin, T., & Sridharan, S. (2004, December). *Confusability of phonemes grouped according to their viseme classes in noisy environments*. Paper presented at the Tenth Australian International Conference on Speech Science & Technology, Sydney, Australia.
- MacDonald, J., & McGurk, H. (1978). Visual influences on speech perception processes. *Perception & Psychophysics*, 24, 253–257. doi:10.3758/BF03206096
- Magnotti, J. F., & Beauchamp, M. S. (2014). The noisy encoding of disparity model of the McGurk effect. *Psychonomic bulletin & review*, 1–9. doi:10.3758/s13423-014-0722-2
- Magnotti, J. F., Ma, W. J., & Beauchamp, M. S. (2013). Causal inference of asynchronous audiovisual speech. *Frontiers in psychology*, 4, 1–10.
- McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, 264, 746–748. doi:10.1038/264746a0
- McKenna Benoit, M., Raij, T., Lin, F.-H., Jääskeläinen, I. P., & Stufflebeam, S. (2010). Primary and multisensory cortical activity is correlated with audiovisual percepts. *Human Brain Mapping*, 31, 526–538.
- Munhall, K. G., Gribble, P., Sacco, L., & Ward, M. (1996). Temporal constraints on the McGurk effect. *Perception & Psychophysics*, 58, 351–362. doi:10.3758/BF03206811
- 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.
- Nath, A. R., Fava, E. E., & Beauchamp, M. S. (2011). Neural correlates of interindividual differences in children's audiovisual speech perception. *Journal of Neuroscience*, 31, 13963–13971.
- Orne, M. (1962). On the social psychology of the psychological experiment: With particular reference to demand characteristics and their implications. *American Psychologist*, 17, 776–783.
- Pearl, D., Yodashkin-Porat, D., Katz, N., Valevski, A., Aizenberg, D., Sigler, M., & Kikinzon, L. (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, 186–192.
- Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision*, 10, 437–442. doi:10.1163/156856897X00366
- Quinto, L., Thompson, W. F., Russo, F. A., & Trehub, S. E. (2010). A comparison of the McGurk effect for spoken and sung syllables. *Attention, Perception, & Psychophysics*, 72, 1450–1454. doi:10.3758/APP.72.6.1450
- R Development Core Team. (2014). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from www.R-project.org
- Sekiyyama, K., Braida, L., Nishino, K., Hayashi, M., & Tuyo, M. (1995). The McGurk effect in Japanese and American perceivers. In K. Elenius & P. Branderud (Eds.), *Proceedings of the XIIIth International Congress of Phonetic Sciences* (Vol. 3, pp. 214–217). Stockholm, Sweden: ICPPhS & Stockholm University.
- Shapiro, S. S., & Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, 52, 591–611.
- Skipper, J., van Wassenhove, V., Nusbaum, H., & Small, S. (2007). Hearing lips and seeing voices: How cortical areas supporting speech production mediate audiovisual speech perception. *Cerebral Cortex*, 17, 2387–2399.
- Stevenson, R. A., Siemann, J. K., Schneider, B. C., Eberly, H. E., Woynaroski, T. G., Camarata, S. M., & Wallace, M. T. (2014). Multisensory temporal integration in autism spectrum disorders. *Journal of Neuroscience*, 34, 691–697. doi:10.1523/JNEUROSCI.3615-13.2014
- Stevenson, R. A., Zemtsov, R. K., & Wallace, M. T. (2012). Individual differences in the multisensory temporal binding window predict susceptibility to audiovisual illusions. *Journal of Experimental Psychology: Human Perception and Performance*, 38, 1517–1529. doi:10.1037/a0027339
- 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.
- Strelnikov, K., Rouger, J., Lagleyre, S., Fraysse, B., Deguine, O., & Barone, P. (2009). Improvement in speech-reading ability by auditory training: Evidence from gender differences in normally hearing, deaf and cochlear implanted subjects. *Neuropsychologia*, 47, 972–979. doi:10.1016/j.neuropsychologia.2008.10.017
- Tremblay, C., Champoux, F., Bacon, B. A., & Théoret, H. (2007a). Evidence for a generic process underlying multisensory integration. *Open Behavioral Science Journal*, 1, 1–4. doi:10.2174/187423000701011000
- Tremblay, C., Champoux, F., Voss, P., Bacon, B. A., Lepore, F., & Théoret, H. (2007b). Speech and non-speech audio-visual illusions: A developmental study. *PLoS One*, 2, e742. doi:10.1371/journal.pone.0000742
- Woynaroski, T. G., Kwakye, L. D., Foss-Feig, J. H., Stevenson, R. A., Stone, W. L., & Wallace, M. T. (2013). Multisensory speech perception in children with autism spectrum disorders. *Journal of Autism and Developmental Disorders*, 43, 2891–2902. doi:10.1007/s10803-013-1836-5