Attention, Perception, & Psychophysics
https://doi.org/10.3758/s13414-022-02517-z

SHORT REPORT

Speech and non-speech measures of audiovisual integration are not correlated

Jonathan M. P. Wilbiks[1] · Violet A. Brown[2] · Julia F. Strand[3]

Accepted: 9 May 2022

Abstract

Many natural events generate both visual and auditory signals, and humans are remarkably adept at integrating information from those sources. However, individuals appear to differ markedly in their ability or propensity to combine what they hear with what they see. Individual differences in audiovisual integration have been established using a range of materials, including speech stimuli (seeing and hearing a talker) and simpler audiovisual stimuli (seeing flashes of light combined with tones). Although there are multiple tasks in the literature that are referred to as "measures of audiovisual integration," the tasks themselves differ widely with respect to both the type of stimuli used (speech versus non-speech) and the nature of the tasks themselves (e.g., some tasks use conflicting auditory and visual stimuli whereas others use congruent stimuli). It is not clear whether these varied tasks are actually measuring the same underlying construct: audiovisual integration. This study tested the relationships among four commonly-used measures of audiovisual integration, two of which use speech stimuli (susceptibility to the McGurk effect and a measure of audiovisual benefit), and two of which use non-speech stimuli (the sound-induced flash illusion and audiovisual integration capacity). We replicated previous work showing large individual differences in each measure but found no significant correlations among any of the measures. These results suggest that tasks that are commonly referred to as measures of audiovisual integration may be tapping into different parts of the same process or different constructs entirely.

Keywords Audiovisual integration · Individual differences · Convergent validity

Given the ubiquity of stimuli that generate simultaneous auditory and visual signals (e.g., objects colliding, faces speaking), it may not be surprising that perceptual systems have evolved to integrate information from these two modalities. Integrating auditory and visual information appears to occur both early (Talsma et al., 2007) and late (Massaro & Cohen, 1995) in the perceptual process and enables speeded responding and improved stimulus identification (Stein & Stanford, 2008). Audiovisual integration requires that perceivers first extract specific pieces of information from both modalities. However, unimodal extraction does not guarantee integration: If a perceiver judges the stimuli to come from different sources, they are less likely to be integrated than if they are assumed to come from the same cause (i.e., the unity assumption; Welch & Warren, 1980). The judgment of whether the stimuli came from a common source involves consideration of several factors, including temporal coincidence, spatial coincidence, and crossmodal congruence (for a comprehensive review, see Koelewijn et al., 2010). Once this probabilistic computation has been completed, the perceiver may integrate the stimuli if appropriate.

Although all forms of audiovisual integration require unimodal extraction and binding of the inputs, the phenomenon has been assessed in many different ways within the literature, including through behavioral tasks (e.g., McGurk & MacDonald, 1976; Shams et al., 2000), neuroimaging paradigms (e.g., Beauchamp et al., 2004; Calvert et al., 2000), and even single-cell recordings (Stein et al., 1976). The stimuli that are employed in research on audiovisual integration also vary, ranging from simple stimuli such as flashes of light and tones (Shams et al., 2000) to naturally occurring complex

✉ Jonathan M. P. Wilbiks
jwilbiks@unb.ca

[1] Department of Psychology, University of New Brunswick, Saint John, NB, Canada

[2] Department of Psychological & Brain Sciences, Washington University in St. Louis, Saint Louis, MO, USA

[3] Department of Psychology, Carleton College, Northfield, MN, USA

Published online: 24 May 2022

stimuli such as audiovisual speech (Sumby & Pollack, 1954). In some cases, stimuli in one modality have been shown to increase the perceptibility of stimuli in another modality (e.g., Sommers et al., 2005; Van der Burg et al., 2008), and in other cases, integration can be measured through perception in one sensory modality being led astray by a stimulus in a different modality (e.g., the sound-induced flash illusion; Shams et al., 2000; the McGurk effect, McGurk & MacDonald, 1976). The fact that audiovisual integration can be demonstrated in so many settings and with so many stimuli is an indication of the robustness of the phenomenon.

A growing body of work suggests that there are also large individual differences in the ability or propensity to integrate visual and auditory input (e.g., Brown et al., 2018; Gurler et al., 2015; Van Engen et al., 2017). Many tasks (e.g., the McGurk effect and the sound-induced flash illusion) that have been used to demonstrate that what we see influences what we hear are also used to quantify individual differences in audiovisual integration. Thus, in addition to being a feature of the perceptual system that is not specific to a single stimulus type or setting, there is now ample evidence that something about the process of audiovisual integration differs across perceivers.

One explanation for the individual differences observed in both speech and non-speech tasks and with both abstract and naturalistic stimuli is that there is an "audiovisual integration ability" on which perceivers systematically differ (note that individuals may differ in multiple aspects of the integration process, such as binding and fusion; Lindborg & Andersen, 2021; Nahorna et al., 2012). That is, audiovisual integration may represent an underlying ability that is not task-specific (see Huang et al., 2012, for an analogous claim in the attention literature). One challenge to assessing whether "audiovisual integration ability" is a unified construct that affects performance on audiovisual tasks across domains of research is that there has been relatively little contact between integration research using speech and non-speech materials. However, some studies that have included multiple tasks purporting to measure audiovisual integration have found that performance on speech and non-speech tasks are indeed related. For example, Stevenson et al. (2012) showed that individual differences in McGurk susceptibility were related to individual differences in integrating non-speech materials. Similarly, Conrey and Pisoni (2006) showed that performance on a non-speech audiovisual synchrony task (identifying whether a visually-presented circle coincided with a tone) was related to audiovisual sentence recognition in noise.[1]

Despite some findings that support the assumption of a common underlying integration ability, we argue that these tasks may be tapping into different features of the process of audiovisual integration (see Odegaard & Shams, 2016). In this paper, we focus on four tasks that have been referred to as "measures of audiovisual integration," including two speech (McGurk and audiovisual benefit) and two non-speech (sound-induced flash illusion and audiovisual integration capacity) tasks. Within each pair of tasks, we included one task in which a stimulus in one modality leads to an illusory perception in the other (speech: McGurk; non-speech: sound-induced flash illusion), as well as one task in which a stimulus in one modality is boosted into greater perceptibility by a stimulus in the other modality (speech: audiovisual benefit; non-speech: audiovisual integration capacity). Below, we describe how these tasks are typically implemented and what they are intended to measure, and then explain why performance on the four tasks might be expected to tap into different features of the integration process and therefore not correlate.

Tasks

McGurk task

The McGurk effect is a classic example of the influence of visual information on auditory perception. Although the implementation varies across studies, McGurk trials typically consist of an auditory stimulus (e.g., "ba") paired with an incongruent visual stimulus (e.g., "ga") that is expected to result in the perception of a third stimulus—a *fusion*—that incorporates components of both the auditory and visual input (e.g., "da"). The proportion of trials on which participants report these fused percepts indicates their susceptibility to the illusion (i.e., the extent to which their perception of the auditory stimulus is affected by the presence of conflicting visual information). Previous work has shown large individual variability in susceptibility to the illusion, with some participants reporting fusion responses on nearly every trial and some consistently reporting the auditory token alone (Basu Mallick et al., 2015; McGurk & MacDonald, 1976).

Audiovisual benefit

Measures of audiovisual benefit assess the extent to which seeing the talker's face increases speech intelligibility relative to hearing them alone (Erber, 1972; Sommers et al., 2005; Sumby & Pollack, 1954). It is typically assessed by presenting participants with speech tokens in noise in both audio-only and audiovisual settings and calculating the relative gains a participant achieves from the addition of the visual signal. The change in performance is typically quantified using the equation $(AV - A)/(1 - A)$, which normalizes improvement from

[1] Note that performance on the non-speech audiovisual synchrony task also predicted accuracy on audio-only sentence recognition, so it is possible that the measure of non-speech synchrony detection used in that experiment was related to a general feature of speech processing common to both audio-only and audiovisual speech rather than integrating auditory and visual information.

seeing the talker relative to the amount each participant could possibly improve (see Grant & Seitz, 1998; Sommers et al., 2005). Audiovisual benefit has been observed for syllables, words, and sentences (Sommers et al., 2005), and at the group level is one of the most robust findings in the speech perception literature.

Sound-induced flash illusion

In the sound-induced flash illusion (Shams et al., 2000), when a single visual stimulus (a white circle) is briefly flashed on the screen along with two rapidly-presented auditory tones, participants often report having seen two flashes. In this task, audiovisual integration is quantified as the proportion of trials on which participants report seeing two flashes; that is, the proportion of trials on which the auditory input influenced what they reported seeing. Although the illusion is typically experienced on approximately half of trials, participants differ in how often they report the illusory percept (e.g., on 20%–78% of trials; Keil et al., 2014).

Audiovisual integration capacity

The audiovisual integration capacity task (Van der Burg et al., 2013) is designed to quantify the number of visual stimuli an individual can successfully integrate with an auditory stimulus. In this task, participants view a field of dots on a computer screen. At regular intervals, a small subset of the dots change color, sometimes coinciding with an aurally presented tone. Participants are asked to hold in memory which dots changed simultaneously with the tone and are then probed about whether one particular dot was among the set that changed. Integration is measured by assessing the number of visual stimuli that are successfully bound with a single auditory stimulus. Initially, raw proportion correct scores are calculated with each number of visual stimuli changing. These scores are then modeled using a least squares method to generate an overall estimate of an individual's capacity for integration. This modeling assumes that if a participant's capacity is greater than the number of dots changing, they will respond correctly. If their capacity is less than the number of dots changing, performance will be predicted by an equation similar to Cowan's (2001) K (more details on modelling are available in Van der Burg et al., 2013). As with the other three tasks described, there is substantial individual variability in performance on the audiovisual integration capacity task, suggesting that individuals differ in their capacity to combine auditory and visual inputs (e.g., Van der Burg et al., 2013).

Comparisons across tasks

Within each of their respective domains, the four tasks described above have been used to measure the construct of "audiovisual integration." But across domains, it seems likely that these tasks may be tapping into different features of audiovisual integration or different steps in the integration process.

One reason that individual differences in performance on speech and non-speech tasks may not correlate involves the complexity of the stimuli. Individual differences in the audiovisual speech tasks may reflect differences in the extent to which the individual can extract phonetic detail from each of the unimodal inputs (i.e., individual differences in hearing ability or lipreading skill; Tye-Murray et al., 2016). Given the simplicity of the stimuli in non-speech tasks (e.g., visually-presented shapes), it is more likely that individual differences on those tasks reflect something about combining what is seen with what is heard, rather than unimodal abilities. That is, in non-speech tasks, the individual must simply detect the presence of a visual stimulus rather than extract meaning from that stimulus. Thus, participants may perform differently on speech and non-speech tasks because even if they both tap into some aspect of the ability to integrate auditory and visual inputs, speech tasks also depend on the ability to extract unimodal phonetic information and infer meaning from the unified percept.

Another major feature on which the tasks differ is whether the information from the auditory and visual channels are congruent. In cases in which the signals do not align (i.e., the McGurk task and the sound-induced flash illusion), this incongruity often leads to an illusory percept. In the speech realm, although audiovisual benefit may depend on individual differences in unimodal extraction ability and integration ability, performance on McGurk tasks may also be influenced by individual differences in the extent to which participants notice the incongruity: Those who detect the incongruity more often may be less likely to integrate the inputs into a unified percept because they are aware that the two signals arise from different sources (Alsius et al., 2017). Tasks using incongruent auditory and visual stimuli therefore may not measure integration alone, but also the ability to detect cross-modal incongruity. In support of this claim, susceptibility to the McGurk effect and audiovisual benefit appear to be uncorrelated (Magnotti et al., 2020; Van Engen et al., 2017; though see Grant & Seitz, 1998). Therefore, despite both being referred to as measures of audiovisual integration, these two speech tasks may be assessing different features of the process (i.e., congruity detection versus extraction and binding of unimodal inputs).

As with the speech tasks, tasks that use non-speech stimuli differ in whether they elicit an illusory perception (the sound-induced flash illusion task) or not (the audiovisual integration capacity task). Therefore, in addition to the ability to extract information from the unimodal inputs and combine them into a unified percept, the sound-induced flash illusion task may assess participants' ability to detect incongruity. This is conceptually similar to the McGurk task, but the incongruity for the sound-induced flash illusion task is temporal, whereas in the McGurk task, it is phonetic.

In addition to audiovisual incongruity, another way in which the non-speech tasks differ is in the information content of the auditory signal. In the audiovisual integration capacity task, the auditory signal serves only as an attentional cue; that is, the tone simply indicates when to focus attention spatially. In the sound-induced flash illusion task, however, the participants must extract information from the auditory signal (i.e., they must determine how many tones were present). Thus, the audiovisual integration capacity task may measure auditory attention in addition to integration, and the sound-induced flash illusion task may also measure the ability to extract meaning from and make decisions about information in multiple modalities simultaneously.

The current study

A relatively small body of work has attempted to directly assess whether performance on multiple tasks that are called "measures of audiovisual integration" are correlated, and there is ample theoretical reason to expect that these tasks may be tapping into different components of the process of integration (see Odegaard & Shams, 2016). However, the limited work that has directly compared speech and non-speech measures of integration has indeed found correlations between them (see Conrey & Pisoni, 2006; Stevenson et al., 2012). The current study aims to test the robustness of those effects and use multiple measures of integration to assess whether the previously observed correlations are task-dependent. Given that no studies to date have assessed multiple measures of audiovisual integration ability including both speech and non-speech materials and both congruent and incongruent stimuli, the current study will use tasks that differ on both of those features.

Assessing whether findings extend across multiple tasks is likely to be informative for at least two reasons. First, when researchers in the speech and non-speech realms describe the mechanisms of integration, they are typically agnostic as to whether their findings are specific to the type of stimuli they are using. Demonstrating differences across tasks may encourage researchers to use greater precision in specifying the mechanisms they are trying to uncover. Second, it has been commonplace to use tasks that employ incongruent stimuli as proxies for naturally occurring audiovisual integration, but it is becoming increasingly clear that the mechanisms underlying incongruent integration may differ from those underlying the integration of congruent stimuli (Alsius et al., 2017). Assessing whether and how the four tasks described above are interrelated will provide a more comprehensive picture of audiovisual integration ability and the bounds of the observed effects.

Method

The study protocol was preregistered on the Open Science Framework. The preregistration document is available at https://osf.io/9rhbs and the raw data, analysis code, and stimuli are available at https://osf.io/e92yr.

Participants

The limited work that has assessed relationships between the measures used in this study has tended to report strong correlations between them: r = −.65 between McGurk susceptibility and sound-induced flash illusion susceptibility (Stevenson et al., 2012), r = .43–.46 between McGurk susceptibility and audiovisual benefit (Grant & Seitz, 1998). To be conservative and account for publication bias (Anderson et al., 2017; Hedges, 1984), we opted to power the study to be able to detect smaller effects. We therefore preregistered a sample size of 150 participants, which can reliably detect an effect of r = .26 with power (1 - β) = .90.[2] This ensures that the study is powered to detect even weaker correlations, such as those our lab has reported between McGurk susceptibility and lipreading: r = .29 (Brown et al., 2018) and r = .32 (Strand et al., 2014). To reach the intended number, we collected data from 158 participants with self-reported normal hearing and normal or corrected-to-normal vision. Seven participants experienced technical difficulties and could not complete the study, and one participant withdrew from the experiment. Participants were recruited from an undergraduate research participant pool at the University of New Brunswick Saint John. Participants were compensated with a bonus point to be used in an undergraduate psychology course. The 150 participants included in the analysis (116 female, 34 male) had a mean age of 21.7 years (SD = 6.4). All recruitment and experimental practices were approved by the Research Ethics Board at the University of New Brunswick Saint John.

[2] Note that we preregistered a sample size of 150 but with an incorrect justification (that it would enable us to detect an effect size of r = .17 rather than r = .26). However, we have collected data from the number of participants originally stated.

McGurk task

Speech stimuli were taken from Brown and Strand (2019). Speech was produced by a female native speaker of American English without a strong regional accent speaking consonants followed by "a." Video files displayed her head and shoulders on a white background. Stimuli consisted of six different tokens of each of four different McGurk stimuli ($A_{ba}V_{ga}$, $A_{ba}V_{fa}$, $A_{ma}V_{ta}$, $A_{pa}V_{ka}$). Each token was repeated four times, resulting in a total of 96 McGurk stimuli. Those stimuli were intermixed with 44 congruent filler stimuli: four tokens each of 11 different syllables ("ba," "da," "fa," "ga," "ka," "ma," "na," "pa," "ta," "ða," and "va") produced by the same talker. After each syllable, participants were asked to report what they perceived by typing it in a text box and pressing enter. Once participants had entered their response, there was a 500 ms interstimulus interval before the next trial.

Audiovisual benefit

The congruent speech task included the same 44 congruent stimuli used in the McGurk task and were intermixed with audio-only presentations of the same 44 stimuli (during which a blank screen was presented). Each stimulus was presented twice for a total of 176 trials. Speech stimuli were presented in a continuous stream of speech-shaped noise, generated in Praat to match the long-term average spectrum of the syllables (Winn, 2018). The signal-to-noise ratio was −8 dB to avoid ceiling-level performance in the audiovisual condition and floor-level performance in the audio-only condition. Participants responded to each trial by typing what they perceived in a text box (with a 500 ms interstimulus interval).

Sound-induced flash illusion

The sound-induced flash illusion task was adapted from the version used by McGovern et al. (2014). A white fixation cross measuring 1° × 1° was presented in the center of a black display. On every trial, a single white circle (diameter = 2°) was visually presented with its center 5° below the fixation cross for 17 ms. The circle was presented with either one or two pure tones at 3.5 kHz frequency, with a duration of 7 ms and an intensity of 95 dB(C).

On single-tone trials, the auditory and visual stimuli were presented with synchronous onset times. These trials were included as fillers to ensure that even the most susceptible participants still would have trials in which they only perceived one flash. On two-tone trials—those that were expected to induce the illusion—one tone was always presented with its onset simultaneous with the onset of the visual stimulus. On "lead" trials, the first tone in a pair occurred before the disc was presented (therefore the second tone was simultaneous with the disc), and on "lag" trials, the first tone was simultaneous and the second tone occurred after the disc was presented. Stimulus onset asynchronies for the lead and lag trials included ± 25, 50, 70, 100, 125, and 150 ms. Each of the 13 timing conditions was presented twice to form a block of 26 trials. Participants completed six blocks, for a total of 156 trials, 144 of which (those with two tones) were included in the final analysis.

Participants were asked to maintain fixation on the cross while attending to the area below it (where the disc would appear). They were asked to attend to both auditory and visual stimuli, and after each trial, participants indicated the number of circles they believed they saw by pressing the left button on a Cedrus RB-540 button box if they saw one, and the right button if they saw two. Given that the visual stimulus was always a single flash, "two flash" responses on two-tone trials indicated that the participants experienced the illusion.

Audiovisual integration capacity

The audiovisual integration capacity task was adapted from a task used in previous studies (e.g., Wilbiks & Beatteay, 2020). In this task, participants were presented with eight dots, 1.5° in diameter, on a grey background. The dots were randomly assigned to be black or white and were arranged in a circle 13° in diameter with a 0.15° fixation dot in the center. After an initial presentation of the dots, between one and four of the dots changed polarity from white to black (or vice versa) repeatedly, 10 times total, at 400 ms per switch. Although previous research (Van der Burg et al., 2013; Wilbiks & Dyson, 2018; Wilbiks et al., 2020) has varied the presentation rate to examine effects of stimulus variation on audiovisual integration capacity, given that we used this task for purposes of comparison to other measures of integration, we used a single presentation rate of 400 ms. Although the number of dots that changed polarity on each switch was constant within a particular trial, the individual dots that changed were not constant. The specific dots that changed polarity were assigned randomly for each switch, with no restriction on which dots could switch.

The penultimate switch was accompanied by an auditory stimulus consisting of a 500 Hz sine tone presented for 60 ms (with 5 ms onset and offset ramps), which was presented at an intensity of approximately 74 dB(C). Participants were instructed to keep track of which dots were changing throughout the trial and were told to note which dot(s) changed in synchrony with the tone. After a 1000 ms retention interval in which only the fixation dot was presented, the final array of dots was displayed again along with a 1° diameter red dot overlaid on one of the dots. Participants were asked to make an unspeeded response as to whether the dot at that location changed at the same time as the tone. They indicated their response by pressing the number 1 on a keyboard if the dot did not change at the same time as the tone, and by pressing the number 2 if the dot did change at the same time as the tone. No feedback was provided, and the subsequent trial began 500–1,500 ms (randomly in increments of 100 ms) after the response was entered. The probe had a validity of 50%, meaning that half the time the red dot indicated the correct answer and half the time it indicated an incorrect answer, and invalid probes were randomly assigned to any invalid location. An experimental block contained 24 trials: three presentations of each number of dots (1, 2, 3, 4) by stimulus validity (valid, invalid) crossing. Each participant completed three blocks, for a total of 72 trials.

Procedure

The four tasks were administered to each participant in the order in which they appear above. All tasks were administered in a dimly lit, quiet room. Visual stimuli were presented on a Dell 2407 WFP monitor at a viewing distance of approximately 57 cm. Auditory stimuli were presented binaurally through Sennheiser HD 280 Pro headphones. Consistent with previous work conducted by our labs, stimulus presentation and data collection were controlled by Superlab (Version 5; Cedrus) for the speech tasks and by Presentation (Version 21.0, Neurobehavioral Systems) for the non-speech tasks.

Results

All data were cleaned, analyzed, and visualized in R version 4.0.4 (R Core Team, 2020) using the following packages: *tidyverse* (Version 1.3.0; Wickham et al., 2019), *broom* (Version 0.7.5; Robinson et al., 2021), *pwr* (Version 1.3.0; Champely, 2020), *psychometric* (Version 2.2; Fletcher, 2015), *ggpubr* (Version 0.4.0; Kassambara, 2020), *data.table* (Version 1.14.0; Dowle & Srinivasan, 2021), *psych* (Version 2.0.12; Revelle, 2021), *Hmisc* (Version 4.5-0; Harrell, 2021), *BayesFactor* (Version 0.9.12-4.2; Morey, 2018), and *bayestestR* (Version 0.8.2; Makowski et al., 2019).

McGurk task

McGurk fusions were defined using the criteria from Brown and Strand (2019). The following responses were scored as fusions: for $A_{ba}V_{ga}$, "da" or "tha"; for $A_{ba}V_{fa}$, "va"; for $A_{ma}V_{ta}$, "na"; and for $A_{pa}V_{ka}$, "ta" or "tha." The rationale for these criteria was that the fusion incorporates elements of both the auditory and visual stimuli. For each participant, McGurk susceptibility was defined as the proportion of McGurk trials on which they reported a fusion response. Some previous work (Stevenson et al., 2012) has normalized McGurk susceptibility relative to accurate identification of audio-only tokens to ensure that any fusion responses are not the result of poor audio quality. We opted against that in this study because the audio-only tokens were prescreened for high intelligibility (Brown & Strand, 2019). Replicating previous work (e.g., Basu Mallick et al., 2015), there was large variability in the extent to which participants were susceptible to the McGurk effect, with fusion scores ranging from 0%–99% (*M* = 55%, *SD* = 29%).

Audiovisual benefit

We first calculated mean accuracy on the audio-only and audiovisual trials separately for each participant. Audio-only syllable identification accuracy ranged from 14%–73% (*M* = 53%, *SD* = 10%) across participants, indicating that the signal-to-noise ratio was successful at keeping performance off the ceiling and floor. Audiovisual syllable identification accuracy ranged from 43%–95% (*M* = 81%, *SD* = 7%) across participants. This average increase of 28 percentage points from the audio-only to the audiovisual condition is consistent with previous research (e.g., Sommers et al., 2005).

Audiovisual benefit scores $(AV - A)/(1 - A)$ ranged from 0.23 to 0.89 (*M* = 0.60, *SD* = 0.12), indicating substantial variability in how much individuals benefited from seeing the talking face. For example, two participants identified 45% of the audio-only syllables correctly, but one of them correctly identified 85% of audiovisual syllables (benefit = 0.73) whereas another identified only 61% of audiovisual syllables (benefit = 0.29).

Sound-induced flash illusion

For each condition in which two tones were presented (ranging from 150 ms lead to 150 ms lag, in 25 ms intervals), we calculated the proportion of trials on which the participant reported perceiving multiple flashes. The preregistered analysis plan called for a calculation of the width of the susceptibility threshold—that is, the range of stimulus onset asynchronies for which an individual participant reported seeing two flashes at least 50% of the time. However, in our data we observed 95 participants who never exceeded 50% susceptibility. We therefore opted to follow an alternative method used in previous research (Hirst et al., 2019; Shams et al., 2002) to calculate the average susceptibility to the illusion across all timepoints for each participant. Individuals varied substantially in their susceptibility to the illusion, with scores ranging from 0%–100% (*M* = 44%, *SD* = 24%).

Audiovisual integration capacity

One participant did not complete the task correctly (they consistently responded prior to the presentation of the red probe dot), so were excluded from analyses involving this task. Estimates of audiovisual integration capacity were calculated by fitting the raw data to a variant of Cowan's K (Cowan, 2001; see Van der Burg et al., 2013; Wilbiks & Beatteay, 2020 for examples of the same approach). This model assumes that an individual's performance (p) is a function of the number of objects to be integrated (n) and their integration capacity (K) and can be modeled with the equation: $p = K/2n + .5$. Thus, if an individual's capacity is equal to or greater than the number of objects to be integrated on a given trial, their performance will be optimal (i.e., if $K \geq n$, $p \approx 1$). An estimate of audiovisual integration capacity is obtained for each participant by finding the value of K that minimizes the root-mean-square error between the raw data (proportion of trials in which they correctly identified whether a given dot changed polarity) and the model predictions. Curve fitting for each participant was performed using four points, each representing the average performance across all trials for a particular dot-changing condition (1–4 changes). The model encountered convergence issues for one participant, so that participant was assigned a K value of 0 (the model with convergence issues had a K estimate of approximately 0). Audiovisual integration capacity ranged from 0.00 to 1.37 ($M = 0.55$, $SD = 0.42$).

Correlational analyses

Correlation coefficients and scatter plots comparing the four measures are shown in Fig. 1. We did not find significant correlations between any of the measures. Given the null findings, we also calculated Bayes Factors to assess the magnitude of the evidence for the null hypothesis (note that this exploratory analysis was not preregistered). Since we were explicitly evaluating the strength of support for the null hypothesis, we calculated the BF_{01} rather than the BF_{10} (which evaluates the

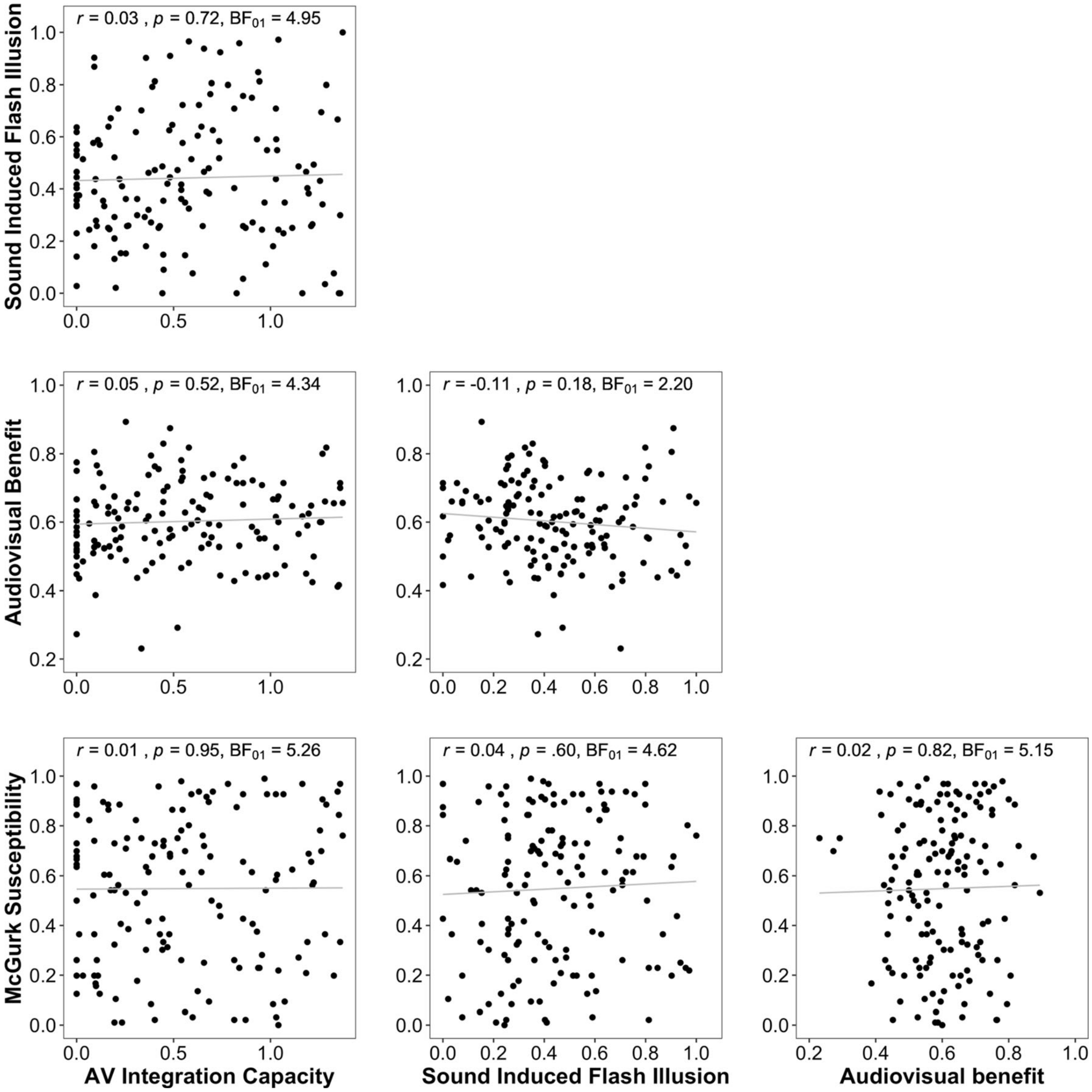

Fig. 1 Correlations among the four tasks, including p values and BF_{01} values

strength of support for the alternative hypothesis). BF_{01} values were all above 2.20, indicating that the observed patterns of data are at least twice as likely to emerge under the null as compared to the alternative hypothesis. The correlation between audiovisual benefit and susceptibility to the sound-induced flash illusion yielded a BF_{01} of 2.20, which is classified as anecdotal evidence for the null hypothesis (Jeffreys, 1998; Lee & Wagenmakers, 2014). However, all other BF_{01} values ranged from 4.34 to 5.26, providing moderate evidence in favor of the null hypothesis.

One possible explanation for a lack of correlations is that the tasks have low reliability. We therefore examined the average split-half reliabilities (using 5,000 random splits) for each of the four tasks—estimated via the "splithalf" function in the *splithalf* package (Parsons, 2020)—as well as average intraclass correlations via the "ICC2.lme" function in the *psychometric* package (Fletcher, 2015). The "splithalf" function estimates Spearman–Brown corrected reliabilities, which provide an indication of internal consistency, and the intraclass correlations indicate the expected correlations between the means of the observed outcomes in each task and the means if the task were to be run again. Split-half reliabilities and intraclass correlations are reported in Table 1. The estimates of split-half reliability and intraclass correlation are identical for each task; this is reassuring given that these two estimation methods are asymptotically equivalent to Cronbach's alpha. We report both for completeness. The reliability estimates for all tasks ranged from 0.80 to 0.98, suggesting that the lack of correlations is not attributable to poor reliability.

Discussion

The purpose of this study was to assess whether tasks that are referred to as "measures of audiovisual integration" appear to be tapping into the same underlying construct. Using a well-powered design, we did not find evidence that any of these measures were correlated; the null hypothesis was 2.20 to 5.26 times more likely than the alternative hypothesis for all task pairs. The lack of correlations does not appear to be attributable to poor reliability or restriction of range, as the tasks demonstrated good split-half reliability, the ranges reported here are comparable to those reported elsewhere, and all tasks showed substantial interindividual variability.

The lack of correlation between McGurk susceptibility and audiovisual benefit is in line with the results of Van Engen et al. (2017), who found no correlation between the two tasks when using syllable stimuli for the McGurk task and sentence stimuli for the audiovisual benefit task. However, syllable identification accuracy is only weakly correlated with sentence identification accuracy for congruent materials (Grant & Seitz, 1998), so the correlation may be more likely to emerge here, in which both the McGurk task and the audiovisual benefit task used the same types of stimuli (syllables). We did not find evidence for this, however, strengthening the claim that McGurk susceptibility and audiovisual benefit are not measuring the same underlying ability (see also Magnotti et al., 2020). Although individual differences in audiovisual benefit and McGurk susceptibility may both be affected by unimodal extraction and integration abilities, the tasks differ in that McGurk susceptibility may also be affected by individual differences in incongruity detection (Strand et al., 2014) or assumptions about causal inference (Magnotti & Beauchamp, 2017). Thus, researchers should exercise caution when drawing generalizations across tasks that use audiovisual benefit and those that use McGurk susceptibility, as these tasks may be tapping into different features of the process of audiovisual speech perception.

As with the speech tasks, we did not find any evidence that susceptibility to the sound-induced flash illusion was related to performance on the audiovisual integration capacity task. Although both have been described as measures of audiovisual integration, the demands they place on participants are quite different. Given that the visual stimuli in the sound-induced flash illusion always occur in one location whereas the stimuli in the audiovisual integration capacity task are spatially distributed, the audiovisual integration capacity task may assess the ability to attend to multiple spatial locations, whereas the sound-induced flash illusion does not. This is in line with research showing that the ability to determine whether auditory and visual stimuli occurred at the same location is

Table 1 Split-half reliability and intraclass correlation coefficient estimates for each of the tasks employed

Task	Split-half reliability [95% confidence interval]	Intraclass correlation coefficient
McGurk susceptibility	0.98 [0.97, 0.98]	0.98
Audiovisual benefit	0.80 [0.75, 0.84]	0.80
Sound-induced flash illusion	0.98 [0.97, 0.98]	0.98
Audiovisual integration capacity	0.85 [0.82, 0.88]	0.85

Note. All reliabilities were calculated on the raw accuracies, not the aggregated scores or difference scores. Reliability of the McGurk susceptibility task was estimated using only incongruent trials. All other tasks included all conditions.

uncorrelated with the ability to detect whether they occurred at the same time (Noel et al., 2018; see also Odegaard & Shams, 2016). Another explanation for the lack of correlation between the audiovisual integration capacity and sound-induced flash illusion tasks is that the latter, like the McGurk task, may be affected by individual differences in incongruity detection or the ability to discern whether multimodal inputs were likely to have arisen from the same source.

Given the lack of correlations within speech tasks and non-speech tasks purported to measure audiovisual integration, it is not surprising that the speech and non-speech measures were also uncorrelated. Our results are inconsistent with one study showing a significant negative correlation between McGurk susceptibility and susceptibility to the sound-induced flash illusion (Stevenson et al., 2012). The cause for this discrepancy is not clear; although there were slight methodological differences between our experiment and the previous study (e.g., the previous study used a single McGurk token whereas we used multiple tokens), there is not a clear mechanistic explanation for why any particular methodological choice would lead to a correlation in Stevenson et al. (2012) but not in our experiment. One difference between the two studies was that Stevenson et al. (2012) included filler trials with multiple flashes in the SIFI task, but we did not. A potential concern with using only single-flash trials is that participants may be biased to report seeing two flashes even when they only saw one in an effort to balance their "one flash" and "two flash" response rates. However, the SIFI susceptibility rate we report here (44% overall) is comparable or even lower than the rates reported in previous work that included filler trials with multiple flashes (47%–69% for the most similar trial types and participant group in Hirst et al., 2019), so it seems unlikely that strategic biases significantly inflated "two flash" response rates in our experiment. Another possible reason for differences between the current findings and those of Stevenson et al. (2012) is the level of statistical power present in each study. The current study was powered to detect effects at the lower bound of the 95% confidence interval around Stevenson's correlation (r = .30–.84) with .90 power (calculated using the *pwr* package in R), so the lack of correlation does not appear to be an issue of statistical power in the current study.

Overall, our findings suggest caution in assuming that multiple tasks are tapping into the underlying construct without providing validity evidence for this claim. Thorndike (1904) described this issue as the *jingle fallacy*, wherein researchers believe that certain tools measure the same construct because they share a name (e.g., "audiovisual integration measures"; see Strand et al., 2021, for other examples of jingle in the speech literature). Given that the tasks appear to be measuring different constructs, we recommend that researchers carefully consider which tasks are most appropriate for the purposes of their study. For example, there is debate about whether individual differences in measures of "integration" reflect differences in the process of integrating information from multiple modalities, or instead reflect differences in unisensory abilities (Strand et al., 2014; Tye-Murray et al., 2016). Thus, if the goal of the study requires assessing integration rather than unisensory abilities, it would be advantageous to identify tasks that are relatively insensitive to individual differences in unisensory ability, or to measure and statistically control for those differences. In addition, incongruent stimuli (such as those used in the McGurk task) may be processed differently than congruent ones (Beauchamp et al., 2010; Calvert et al., 2000) because they may require detecting and resolving conflict between the auditory and visual inputs or trying to assign poor exemplars to discrete phoneme categories (see Brown & Strand, 2019). Thus, if cross-modality incongruity is not central to the research question, then researchers should consider using a task with congruent audiovisual stimuli to measure the construct of interest. Finally, we recommend that researchers conduct additional validation work to attempt to better understand what our measures are actually tapping into. Doing so would contribute to an ongoing effort to improve measurement practices in all areas of psychology (Flake & Fried, 2020).

Open practices statement The experiment, data analysis plans, and hypotheses were preregistered prior to data collection. All raw data, analysis code, stimuli, and the preregistration document are available on the Open Science Framework (https://osf.io/e92yr).

Acknowledgements This work was supported by a National Institutes of Health grant via the National Institute on Deafness and Communication Disorders awarded to Julia Strand (R15-DC018114), a Discovery Grant from the Natural Sciences and Engineering Research Council of Canada awarded to Jonathan Wilbiks (2020-04291), and the National Science Foundation through a Graduate Research Fellowship awarded to Violet Brown (DGE-1745038). Many thanks to Kimberly Mazerolle and Michaela Ritchie for their assistance with data collection.

References

Alsius, A., Paré, M., & Munhall, K. G. (2017). Forty years after hearing lips and seeing voices: The McGurk effect revisited. *Multisensory Research, 31*(1/2), 111–144.

Anderson, S. F., Kelley, K., & Maxwell, S. E. (2017). Sample-size planning for more accurate statistical power: A method adjusting sample effect sizes for publication bias and uncertainty. *Psychological Science, 28*(11), 1547–1562.

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.

Beauchamp, M. S., Lee, K. E., Argall, B. D., & Martin, A. (2004). Integration of auditory and visual information about objects in superior temporal sulcus. *Neuron, 41*(5), 809–823.

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. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience, 30*(7), 2414–2417.

Brown, V. A., & Strand, J. F. (2019). “Paying” attention to audiovisual speech: Do incongruent stimuli incur greater costs? *Attention, Perception, & Psychophysics, 81*(6), 1743–1756.

Brown, V. A., Hedayati, M., Zanger, A., Mayn, S., Ray, L., Dillman-Hasso, N., & Strand, J. F. (2018). What accounts for individual differences in susceptibility to the McGurk effect? *PLOS ONE, 13*(11). https://doi.org/10.1371/journal.pone.0207160

Calvert, G. A., Campbell, R., & Brammer, M. J. (2000). Evidence from functional magnetic resonance imaging of crossmodal binding in the human heteromodal cortex. *Current Biology: CB, 10*(11), 649–657.

Champely, S. (2020). *Package “pwr”* (Version 1.3-0) [Computer software]. https://cran.r-project.org/web/packages/pwr/pwr.pdf

Conrey, B., & Pisoni, D. B. (2006). Auditory-visual speech perception and synchrony detection for speech and nonspeech signals. *The Journal of the Acoustical Society of America, 119*(6), 4065–4073.

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *The Behavioral and Brain Sciences, 24*(1), 87–114.

Dowle, M., & Srinivasan, A. (2021). *data.table* (Version 1.14.0) [Computer software]. Comprehensive R Archive Network (CRAN). https://cran.r-project.org/web/packages/data.table/index.html

Erber, N. P. (1972). Auditory, visual, and auditory-visual recognition of consonants by children with normal and impaired hearing. *Journal of Speech and Hearing Research, 15*(2), 413–422.

Flake, J. K., & Fried, E. I. (2020). Measurement schmeasurement: Questionable measurement practices and how to avoid them. *Advances in Methods and Practices in Psychological Science*. https://doi.org/10.31234/osf.io/hs7wm

Fletcher, T. D. (2015). *psychometric* (Version 2.2) [Computer software]. https://cran.r-project.org/web/packages/psychometric

Grant, K. W., & Seitz, P. F. (1998). Measures of auditory–visual integration in nonsense syllables and sentences. *Journal of the Acoustical Society of America, 104*(4), 2438–2450.

Gurler, D., Doyle, N., Walker, E., Magnotti, J., & Beauchamp, M. (2015). A link between individual differences in multisensory speech perception and eye movements. *Attention, Perception, & Psychophysics, 77*(4), 1333–1341.

Harrell, F. E. (2021). *Hmisc: Harrell miscellaneous* (Version 4.5-0) [Computer software]. https://cran.r-project.org/web/packages/Hmisc/Hmisc.pdf

Hedges, L. V. (1984). Estimation of effect size under nonrandom sampling: The effects of censoring studies yielding statistically insignificant mean differences. *Journal of Educational and Behavioral Statistics, 9*, 61–85.

Hirst, R. J., Setti, A., Kenny, R. A., & Newell, F. N. (2019). Age-related sensory decline mediates the sound-induced flash illusion: Evidence for reliability weighting models of multisensory perception. *Scientific Reports, 9*(1), 1–12.

Huang, L., Mo, L., & Li, Y. (2012). Measuring the interrelations among multiple paradigms of visual attention: an individual differences approach. *Journal of Experimental Psychology: Human Perception and Performance, 38*(2), 414–428.

Jeffreys, H. (1998). *The theory of probability*. Oxford University Press.

Kassambara, A. (2020). *ggpubr* (Version 0.4.0) [Computer software]. https://CRAN.R-project.org/package=ggpubr

Keil, J., Müller, N., Hartmann, T., & Weisz, N. (2014). Prestimulus beta power and phase synchrony influence the sound-induced flash illusion. *Cerebral Cortex, 24*(5), 1278–1288.

Koelewijn, T., Bronkhorst, A., & Theeuwes, J. (2010). Attention and the multiple stages of multisensory integration: A review of audiovisual studies. *Acta Psychologica, 134*(3), 372–384.

Lee, M. D., & Wagenmakers, E.-J. (2014). *Bayesian Cognitive modeling: A practical course*. Cambridge University Press.

Lindborg, A., & Andersen, T. S. (2021). Bayesian binding and fusion models explain illusion and enhancement effects in audiovisual speech perception. *PLOS ONE, 16*(2), Article e0246986.

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), Article 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: A Journal Devoted to the Study of the Nervous System and Behavior, 133*, 371–383.

Makowski, D., Ben-Shachar, M., & Lüdecke, D. (2019). BayestestR: Describing effects and their uncertainty, existence and significance within the Bayesian framework. *Journal of Open Source Software, 4*(40), 1541.

Massaro, D. W., & Cohen, M. M. (1995). Perceiving talking faces. *Current Directions in Psychological Science, 4*(4), 104–109.

McGovern, D. P., Roudaia, E., Stapleton, J., McGinnity, T. M., & Newell, F. N. (2014). The sound-induced flash illusion reveals dissociable age-related effects in multisensory integration. *Frontiers in Aging Neuroscience, 6*, 250.

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature, 264*. https://doi.org/10.1038/264746a0

Morey, R. (2018). *BayesFactor: Computation of Bayes factors for common designs* (Version 0.9.12-4.2) [Computer software]. https://cran.r-project.org/package=BayesFactor

Nahorna, O., Berthommier, F., & Schwartz, J.-L. (2012). Binding and unbinding the auditory and visual streams in the McGurk effect. *The Journal of the Acoustical Society of America, 132*(2), 1061–1077.

Noel, J.-P., Modi, K., Wallace, M. T., & Van der Stoep, N. (2018). Audiovisual integration in depth: Multisensory binding and gain as a function of distance. *Experimental Brain Research: Experimentelle Hirnforschung. Experimentation Cerebrale, 236*(7), 1939–1951.

Odegaard, B., & Shams, L. (2016). The brain’s tendency to bind audiovisual signals is stable but not general. *Psychological Science, 27*(4), 583–591.

Parsons, S. (2020). *splithalf: Robust estimates of split half reliability* (Version 0.7.2) [Computer software]. https://doi.org/10.6084/m9.figshare.11956746.v4

R Core Team. (2020). *R: A language and environment for statistical computing* [Computer software]. R Foundation for Statistical Computing. http://www.R-project.org/

Revelle, W. (2021). *psych: Procedures for personality and psychological research* (Version 2.0.12) [Computer software]. https://cran.r-project.org/web/packages/psych/index.html

Robinson, D., Hayes, A., & Couch, S. (2021). *broom* (Version 0.7.5) [Computer software]. https://www.rdocumentation.org/packages/broom/versions/0.7.5

Shams, L., Kamitani, Y., & Shimojo, S. (2000). Illusions. What you see is what you hear. *Nature, 408*(6814), 788.

Shams, L., Kamitani, Y., & Shimojo, S. (2002). Visual illusion induced by sound. *Brain Research. Cognitive Brain Research, 14*(1), 147–152.

Sommers, M. S., Tye-Murray, N., & Spehar, B. (2005). Auditory-visual speech perception and auditory-visual enhancement in normal-hearing younger and older adults. *Ear and Hearing, 26*(3), 263–275.

Stein, B. E., & Stanford, T. R. (2008). Multisensory integration: Current issues from the perspective of the single neuron. *Nature Reviews Neuroscience, 9*(5), 255–266. https://doi.org/10.1038/nrn2377

Stein, B. E., Magalhaes-Castro, B., & Kruger, L. (1976). Relationship between visual and tactile representations in cat superior colliculus. *Journal of Neurophysiology, 39*(2), 401–419.

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*(6), 1517–1529.

Strand, J. F., 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: JSLHR, 57*(6), 2322–2331.

Strand, J. F., Ray, L., Dillman-Hasso, N. H., Villanueva, J., & Brown, V. A. (2021). Understanding speech amid the jingle and jangle: Recommendations for improving measurement practices in listening effort research. *Auditory Perception & Cognition*, 1–20.

Sumby, W. H., & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *The Journal of the Acoustical Society of America, 26*(2), 212–215.

Talsma, D., Doty, T. J., & Woldorff, M. G. (2007). Selective attention and audiovisual integration: is attending to both modalities a prerequisite for early integration?. *Cerebral Cortex, 17*(3), 679–690.

Thorndike, E. L. (1904). *An introduction to the theory of mental and social measurements*. Columbia University, Teacher's College.

Tye-Murray, N., Spehar, B., Myerson, J., Hale, S., & Sommers, M. S. (2016). Lipreading and audiovisual speech recognition across the adult lifespan: Implications for audiovisual integration. *Psychology and Aging, 31*(4), 380–389.

Van der Burg, E., Olivers, C. N., Bronkhorst, A. W., & Theeuwes, J. (2008). Pip and pop: nonspatial auditory signals improve spatial visual search. *Journal of Experimental Psychology: Human Perception and Performance, 34*(5), 1053.

Van der Burg, E., Awh, E., & Olivers, C. N. L. (2013). The capacity of audiovisual integration is limited to one item. *Psychological Science, 24*(3), 345–351.

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.

Welch, R. B., & Warren, D. H. (1980). Immediate perceptual response to intersensory discrepancy. *Psychological Bulletin, 88*(3), 638–667.

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T., Miller, E., Bache, S., Müller, K., Ooms, J., Robinson, D., Seidel, D., Spinu, V., … Yutani, H. (2019). Welcome to the Tidyverse. *Journal of Open Source Software*, *4*(43), 1686.

Wilbiks, J. M., & Dyson, B. J. (2018). The contribution of perceptual factors and training on varying audiovisual integration capacity. *Journal of Experimental Psychology: Human Perception and Performance, 44*(6), 871.

Wilbiks, J. M. P., & Beatteay, A. (2020). Individual differences in multiple object tracking, attentional cueing, and age account for variability in the capacity of audiovisual integration. *Attention, Perception, & Psychophysics, 82*(7), 3521–3543.

Wilbiks, J. M., Pavilanis, A. D., & Rioux, D. M. (2020). Audiovisual integration capacity modulates as a function of illusory visual contours, visual display circumference, and sound type. *Attention, Perception, & Psychophysics, 82*(4), 1971–1986.

Winn, M. B. (2018). *Praat script for creating speech-shaped noise* (Version 12) [Computer software]. http://www.mattwinn.com/praat.html