Bi-directional audiovisual influences on temporal modulation discrimination

Leonard Varghese, Samuel R. Mathias, Seth Bensussen, Kenny Chou, Hannah R. Goldberg, Yile Sun, Robert Sekuler, and Barbara G. Shinn-Cunningham

Citation: The Journal of the Acoustical Society of America **141**, 2474 (2017); doi: 10.1121/1.4979470
View online: http://dx.doi.org/10.1121/1.4979470
View Table of Contents: http://asa.scitation.org/toc/jas/141/4
Published by the Acoustical Society of America

Bi-directional audiovisual influences on temporal modulation discrimination

Leonard Varghese,[1,a)] Samuel R. Mathias,[2] Seth Bensussen,[1] Kenny Chou,[1] Hannah R. Goldberg,[3] Yile Sun,[4] Robert Sekuler,[5] and Barbara G. Shinn-Cunningham[1]
[1]*Department of Biomedical Engineering, Boston University, Boston, Massachusetts 02215, USA*
[2]*Neurocognition, Neurocomputation and Neurogenetics (n3) Division, Yale University School of Medicine, New Haven, Connecticut 06511, USA*
[3]*Center for Computational Neuroscience and Neural Technology, Boston University, Boston, Massachusetts 02215, USA*
[4]*Department of Psychology, Brandeis University, Waltham, Massachusetts 02453, USA*
[5]*Volen Center for Complex Systems, Brandeis University, Waltham, Massachusetts 02453, USA*

(Received 10 January 2017; revised 22 February 2017; accepted 16 March 2017; published online 10 April 2017)

Cross-modal interactions of auditory and visual temporal modulation were examined in a game-like experimental framework. Participants observed an audiovisual stimulus (an animated, sound-emitting fish) whose sound intensity and/or visual size oscillated sinusoidally at either 6 or 7 Hz. Participants made speeded judgments about the modulation rate in either the auditory or visual modality while doing their best to ignore information from the other modality. Modulation rate in the task-irrelevant modality matched the modulation rate in the task-relevant modality (congruent conditions), was at the other rate (incongruent conditions), or had no modulation (unmodulated conditions). Both performance accuracy and parameter estimates from drift-diffusion decision modeling indicated that (1) the presence of temporal modulation in both modalities, regardless of whether modulations were matched or mismatched in rate, resulted in audiovisual interactions; (2) congruence in audiovisual temporal modulation resulted in more reliable information processing; and (3) the effects of congruence appeared to be stronger when judging visual modulation rates (i.e., audition influencing vision), than when judging auditory modulation rates (i.e., vision influencing audition). The results demonstrate that audiovisual interactions from temporal modulations are bi-directional in nature, but with potential asymmetries in the size of the effect in each direction.
 [http://dx.doi.org/10.1121/1.4979470]

[AKCL] Pages: 2474–2488

I. INTRODUCTION

Inputs to different senses that share common properties interact with one another and shape perception. These interactions can be automatic and obligatory, occurring even when attention is directed away from one of the sensory modalities (Molholm *et al.*, 2007). Temporal properties of sensory inputs can influence the strength and nature of cross-sensory interactions (Spence, 2011). For example, when auditory and visual stimuli turn on and off together (Spence and Squire, 2003; Kubovy and Yu, 2012) or (more generally) have correlated amplitudes (Parise *et al.*, 2012; Denison *et al.*, 2013; Maddox *et al.*, 2015), the inputs are likely to be perceived as originating from a single source (fuse into one perceptual object). This in turn can make it difficult to access information about a feature in one sensory modality, independent of the information in the other modality, even when an observer attempts to ignore the second modality. Alternatively, audio-visual interactions could arise via more general cognitive biasing mechanisms, even if audio and visual information are not necessarily bound at the perceptual level (Bizley *et al.*, 2016).

Two recent studies of multisensory temporal processing examined the influence of dynamic auditory stimuli on judgments of visual modulation rates in a multisensory scene offered by a video game (*Fish Police!*; Goldberg *et al.*, 2015; Sun *et al.*, 2016). Players were asked to judge the rate at which a visual stimulus (a computer-generated fish) was modulated in size while it moved across the display and emitted an intensity-modulated sound. The modulation rate of the sound either matched or was incongruent with ("mismatched") the visual modulation rate. Results demonstrate the robustness of "auditory-driving" effects (Gebhard and Mowbray, 1959; Shipley, 1964; Welch and Warren, 1980): auditory information can drive the perception of visual temporal properties. Specifically, the studies found more errors in identifying the visual modulation rate when visual and auditory inputs were mismatched, even though participants were instructed to ignore the sounds (i.e., to focus cross-modal attention; see Spence and Driver, 1997). We were motivated to use a video-game like environment to study multi-sensory processing by a desire to make the tasks more engaging for participants than typical psychological/psychophysical tasks. Indeed, gamifying psychophysical tasks may improve participant motivation to try their best on the task (Washburn, 2003; Miranda and Palmer, 2014). The current study uses the same *Fish Police!* environment for

[a)]Electronic mail: lennyv@bu.edu

similar reasons. However, the current study differs from the previous studies using the same game environment in its treatment of the response data and by expanding on the tasks that participants were asked to perform.

The first difference between the current study and the previous studies using *Fish Police!* is in its treatment of reaction time (RT) data. Audiovisual interactions are known to affect RT on speeded response tasks (Marks, 1987). However, analysis of RT data is often difficult in the presence of differences in accuracy due to speed-accuracy trade-offs (Wickelgren, 1977; Heitz, 2014) or distributional changes in RT across conditions (Speckman *et al.*, 2008). Additionally, consideration of reaction times only on "correct" trials, as in the previous *Fish Police!* studies, may offer an incomplete picture of how audiovisual interactions shape perception. A class of decision-making models known as "sequential sampling models" can be simultaneously fit to the RT distributions of both correct and error responses and thus naturally deal with such issues (Forstmann *et al.*, 2016). The drift-diffusion model (DDM; Ratcliff, 1978; Ratcliff and Rouder, 1998; Ratcliff and McKoon, 2008) is the archetype of this model class. DDMs include parameters meant to isolate and separately represent the quality of stimulus evidence ("drift rate"; higher quality evidence leads to faster decisions), decision caution ("decision threshold"; indicative of observer caution), *a priori* response bias ("bias"), and the combination of stimulus encoding and motor response times ("non-decision time"). DDMs can thus provide insights into response data that are not easily achieved using separate, model-free analyses (Wagenmakers, 2009; Mathias, 2016). Here, response data were considered using a conventional analysis method considering correct/incorrect responses only (mixed-effect logistic regression), as well as using DDM. Since the accuracy data provided to each model was identical, we expected the conclusions drawn from each model to be qualitatively similar. However, we predicted that the DDM would provide additional insight into task performance due to the incorporation of RT into the response model.

The present study also differs from the previous *Fish Police!* experiments by examining whether the cross-modal interactions affected perceived modulation that were observed studies are "bi-directional": whether temporal modulations of visual features influence auditory modulation rate judgments just as auditory intensity modulations influence visual modulation rate judgments. While visual information can affect spatial features of an auditory percept (Soto-Faraco *et al.*, 2002; Soto-Faraco *et al.*, 2004), visual modulations may not have strong effects on judgments of auditory modulation *rates*, since perception of the temporal properties of a scene tend to be dominated by auditory information (Welch and Warren, 1980; Recanzone, 2002; Shams *et al.*, 2002; Michalka *et al.*, 2015). We hypothesized that visual modulations influence the accuracy of judgments of auditory modulation rate, just as auditory modulations influence judgments of visual modulation rates, but that these effects are weaker than those of auditory modulations on visual modulation rate.

II. METHODS

A. Experimental framework and game environment

Participants completed two tasks (an auditory task and a visual task) within the framework of the *Fish Police!* open-source game (Hickey, 2013). Participants completed the two tasks in different sessions, completed on different days. The order in which the sessions were completed was counterbalanced across participants.

On each trial within the game, a single computer-generated image of a fish appeared on either the left or right side of the game window and "swam" across the window on a randomized two-dimensional path (Fig. 1). The fish was always accompanied by a harmonic complex tone with an interaural time difference of 600 μs, leading in the ear matching the fish's starting side (note that the spatial location of the sound accompanying the fish was fixed, and not updated as the fish moved along the screen). The location of the sound and the originating side of the fish were never mismatched (e.g., it was not possible to have the fish image originate on the left and the tone originate on the right). The onsets of the fish image and tone were simultaneous. Throughout the experiment, an additional continuous background sound, meant to evoke the sound of water rushing in a stream, was presented at a low level (thus overlapping in time with the sound associated with each fish); pilot experiments indicated that this additional sound did not result in any perceptual masking of the tones.

The visual and auditory stimuli were either modulated at 6 or 7 Hz, or were not modulated, depending on the task and experimental condition. Visual modulation was achieved by varying the size of the fish sinusoidally in the vertical dimension between 70% and 125% of its original size. Modulation of the auditory stimulus was achieved by sinusoidally varying the amplitude envelope with 50% modulation depth. These rates and modulation depths were kept the same throughout all parts of the experiment.

On each trial in the visual task, participants judged whether the size of the fish was changing at a "fast" (7 Hz) or "slow" (6 Hz) rate. The tone accompanying the fish could be modulated at the same rate (congruent condition), modulated at the other rate (incongruent condition), or have no modulation (unmodulated condition); participants were informed that any auditory modulations were task-irrelevant and that they should ignore them. In the auditory task, participants made fast/slow judgments about the tone that accompanied the fish on each trial. In this task, the size of the fish could be congruent, incongruent, or unmodulated. In the auditory task, participants were informed that the visual modulations were task-irrelevant. In both the visual and auditory tasks, observers were also asked to identify the starting location of the fish on each trial (left/right). Specifically, participants pushed one of four "shoulder" (top) buttons on a Logitech F310 gamepad (see Fig. 1) to indicate the perceived modulation rate of the fish in the attended sensory modality (slow or fast, corresponding to the rear or front buttons, respectively), and the location of the fish at the start of each trial (left or right, corresponding to the left-side and right-side buttons).

After entering a response for a given trial, the fish would disappear from view. Feedback was provided to participants

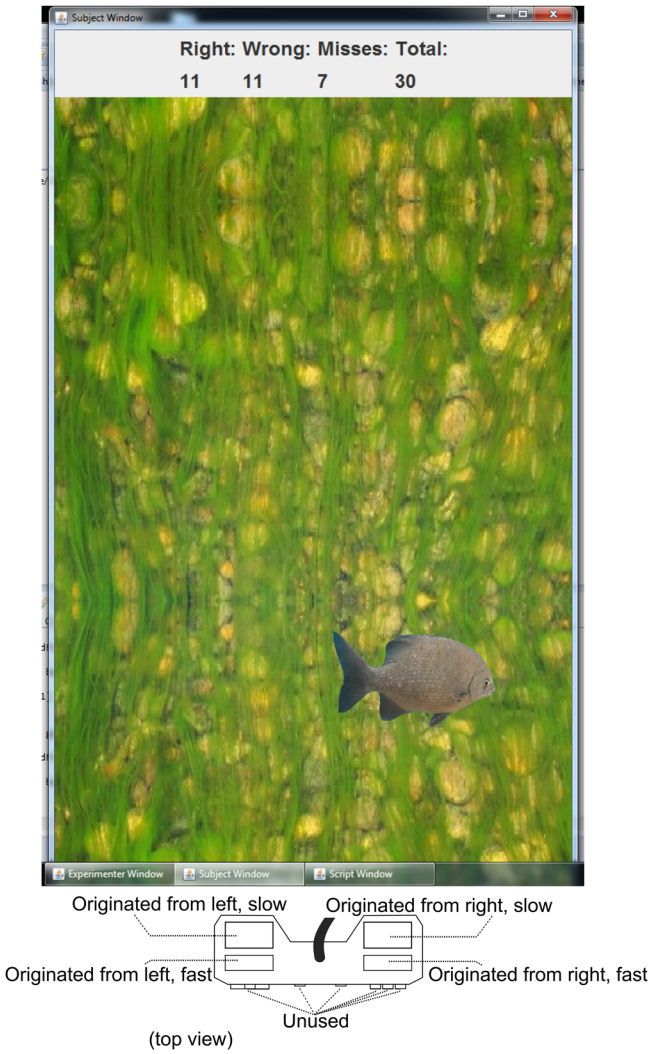

FIG. 1. (Color online) Screenshot from *Fish Police!* as utilized in the current experiment. On each trial, a fish (visual stimulus) swam across the screen in a randomized two-dimensional path. The fish was always accompanied by a tone (auditory stimulus), presented via headphones with an ITD leading to the side matching the side of the screen from which the fish originated. The size of the fish and/or the tones were modulated at 6 or 7 Hz in size (visual stimulus) or amplitude (tone amplitude) to create the experimental conditions (see text). Participants were asked to discriminate the rate at which the visual stimulus or the auditory stimulus was modulated via button presses on a gamepad (see illustration at bottom; top view of gamepad shown). A running tally of correct responses, incorrect responses, and missed (no-response) trials was displayed at the top of the game window, updated after each event.

both in the form of a running tally of correct, incorrect, and missed responses at the top of the screen (see Fig. 1) and a high pitched "ding" or a low-pitched "buzz" after each correct and incorrect responses, respectively. Each trial within the game was timed; the next trial would automatically start after either 2 or 2.5 s at random. A response was registered as "missed" if the participant failed to press one of the four buttons before a new fish appeared. Since the primary experimental goal was investigation of audiovisual congruence on modulation rate judgments, and because errors in location judgment turned out to be exceedingly rare, trials in which a location error was made were not considered in any subsequent analyses (further details below).

Each session began with a series of practice blocks consisting of 20 trials containing congruent, incongruent, and unmodulated trials, randomly intermingled. Practice blocks were repeated until participants scored at least 70% correct on block. After reaching this criterion, subjects completed 6 test blocks, each comprising 150 trials (50 each of congruent, incongruent, and unmodulated trials in random order). Training effects over the course of the six blocks were found to be relatively small: the odds of responding correctly increased by 5% per block for the visual task, and by 9% per block for auditory task (see supplemental Fig. 1).[1] Since training effects were not the primary focus of the study, the effect of block was not considered further.

In summary, the experiment had a $2 \times 2 \times 3$ factorial design: the first factor was the task (visual or auditory); the second factor was the modulation rate of the stimulus in the target modality (slow or fast); and the third factor was the modulation of the task-irrelevant stimuli (congruent, incongruent, or unmodulated). Excluding practice trials (which were not analyzed), participants completed a total of 150 trials per combination of task, congruence, and modulation rate.

B. Testing environment

Participants were tested individually while seated in a dark sound-attenuating booth (Industrial Acoustics, Brooklyn, NY) containing a computer LCD display (Dell 2707WFP, operating at a resolution of 1920×1080), headphones (Sennheiser HD600, Sennheiser electronic GmbH & Co., Hanover, Germany) connected via a USB sound card (MOTU Microbook, MOTU Inc., Cambridge, MA), and the USB gamepad. Distance from the monitor, luminance, and sound presentation level were not controlled precisely; participants sat a comfortable distance away from the monitor (about 40 cm) and sounds were presented at an audible, yet comfortable, level. The game window (playable area: 635 pixels wide, 968 pixels high, with the bottom 30 pixels partially obscured by the translucent Windows OS Taskbar; see Fig. 1) was placed so that it occupied roughly the middle 1/3 of the screen. This positioning of the game window and participant distance from the monitor led each visual stimulus to subtend a visual angle of approximately 7.6°. Although participants seemed likely to track the visual stimulus as it moved across the screen when performing the visual task, such a strategy is potentially detrimental to performance during the auditory task. For this reason, we wished to minimize the possibility that participants averted their gaze from the game window and thus shut out visual information when they performed the auditory task. For this reason, participants were instructed to keep their eyes open, direct their gaze toward the monitor at all times, and not shift their gaze to the sides of the screen (i.e., they were instructed to look at the game window, and not the Windows Desktop on either side of the game window). The experimenter monitored cooperation with these instructions via an infrared camera positioned inside the booth, facing the participant, which was connected to a display outside the booth. The experimenter was able to monitor participants' performance in real

time from outside the booth using a display that mirrored the one on which fish were displayed to participants.

C. Participants

Thirteen participants (7 male, 6 female; age 20–27 year old) were recruited from the Boston University community and were compensated $12.50/h for taking part in the study. All participants signed informed consent documents approved by the Boston University Charles River Campus Institutional Review Board. Participants were screened to ensure that they had normal hearing at standard audiometric frequencies between 200 and 8000 Hz. All participants reported having normal or corrected-to-normal visual acuity. Participants completed each session in less than 1.5 h, including any elective breaks.

D. Data analysis

1. Game log parsing and data pre-processing

Game log files were parsed using MATLAB (v2013a), and subsequent analysis was performed using custom scripts written in Python (version 3) and R (version 3.3.1). A few game logs became corrupted and resulted in some data (2 trials) being dropped from the analysis at the parsing stage. "Missed" trials and trials immediately following "missed" trials (114 trials) were also removed; the latter were removed to ensure correct scoring, because responses on these trials may have been intended as responses to the preceding "missed" trials.

Given the design of the experiment, participants could make three different kinds of errors: they could incorrectly report the modulation rate of the target stimulus while correctly reporting the initial location of the stimulus; they could incorrectly report the location while correctly reporting the modulation rate; or they could incorrectly report the target modulation rate and the location. Incorporation of the location judgments into the task was originally intended to inform another line of investigation on audiovisual spatial judgments. However, the rarity of location judgment errors (11 trials in total across all participants) meant that, in practice, the experiment reduced to a speeded binary decision task on each trial. Therefore, to simplify analysis, trials with location errors were dropped from analysis, and modulation judgments were collapsed across locations. This made the accuracy data suitable for consideration using logistic regression, and the reaction times suitable for analysis with DDM. Since abnormally short reaction times are known to interfere with estimation of DDM parameters, trials with reaction times less than 500 ms (19 trials) were also excluded from the analysis. Ultimately, of the 23 400 experimental trials conducted, 23 254 trials were retained for analysis.

2. DDM fitting

In several respects, applying the DDM to data from speeded two-choice decision tasks is similar to applying signal detection theory (SDT; Green and Swets, 1966). Like SDT, the DDM represents a mathematical implementation of an optimal decision process given sensory inputs contaminated by internal noise (Bogacz *et al.*, 2006). Also like SDT, the DDM contains several latent parameters with distinct psychological interpretations (Wagenmakers, 2009; Mathias, 2016). However, unlike SDT, the DDM is fitted to the RT distributions of correct and error responses, rather than just counts of responses, and contains more latent parameters than SDT. Specifically, the DDM contains four main parameters (Fig. 2), and three optional parameters. The main parameters are (1) the drift rate (v), representing the quality of stimulus evidence available on each trial and the resulting rate at which information leading up to a decision is accumulated; (2) the distance between the decision boundaries (a), representing the degree of certainty the observer requires before responding; (3) the starting point of the diffusion process (z), representing *a priori* bias toward making one response or the other; and (4) the non-decision time (t), representing factors influencing reaction times that are not directly related to the decision process, such as time required for sensory encoding, motor preparation, and response execution (Ratcliff and McKoon, 2008; Voss *et al.*, 2013). The optional parameters reflect trial-to-trial variability in drift rate, bias, and non-decision times (denoted sv, sz, and st, respectively); these are necessary to explain various phenomena observed in real data, such as faster/slower error responses, but prohibitively large numbers of trials are usually needed to obtain meaningful estimates at the level of individual participants (see Wiecki *et al.*, 2013).

DDMs were fitted to the data using the HDDM Python package, version 0.6 (Wiecki *et al.*, 2013). HDDM performs Bayesian hierarchal estimation of DDM parameters via Markov Chain Monte Carlo (MCMC) sampling from the joint posterior distribution using PyMC (version 2.3.3) (Patil *et al.*, 2010). In this software package, individual-participant parameters are estimated hierarchically, with mildly

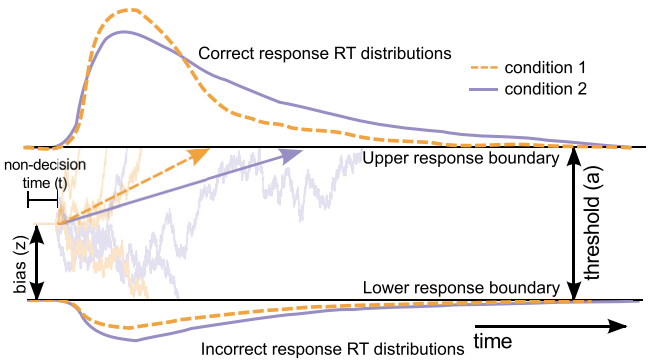

FIG. 2. (Color online) Example of drift-diffusion model (DDM) for a two-condition task. Dashed lines correspond to one set of responses and distributions, solid lines correspond to another set. On each trial, the process leading to a decision is modeled as a random walk between the two boundaries; after a fixed time (non-decision time t, which represents stimulus encoding time), the random walk starts at point z (bias). When the random walk hits either decision boundary (one located at position 0, the other located a arbitrary units away), a response is executed. Differences in how often and when the walk hits one of the decision boundaries leads to observed differences in performance and reaction times, respectively, in a decision-making task. The drift rate (v) can be thought of as similar to a vector sum or average of the random walks; its direction indicates a favored decision, and the magnitude is indicative of the speed with which that decision is made. Adapted from Wiecki *et al.* (2013) with permission.

informative priors based on previously reported DDM parameter estimates placed on the group distributions (Matzke and Wagenmakers, 2009). The individual-participant parameters both influence and are influenced by the estimates of the group parameters, which greatly improves parameter estimation relative to traditional maximum likelihood (Ratcliff and Childers, 2015). Independent models were fit to data from the visual task and from the auditory task, which was justified based on the lack of correlation between performance on audio and visual tasks (see Sec. III).

Estimates of the drift rate (v), decision threshold (a), bias (z), and non-decision time (t) were permitted to vary as a function of the congruence of the auditory and visual stimuli. The effects of congruence were entered into the model using a "treatment-coding" scheme, with the unmodulated condition serving as the baseline (intercept). The consequence of this coding scheme is the same as the consequence of using treatment coding in a linear model: parameters associated with the congruent and incongruent conditions were estimated as changes relative to the values associated with the unmodulated parameter, rather than to the overall mean or another reference point (Faraway, 2014). Distributions of the mean and the standard deviation (pooled across congruence conditions) of each of the four parameters were incorporated into the hierarchical model. Inter-trial variability parameters (sv, st, and sz) were included in the models, but were only estimated at the group level, and not as a function of congruence. For simplicity and to reduce computational demands, the effects of stimulus rate (fast or slow) on any parameter were not considered; instead, stimulus rate was implicitly encoded into the model by associating the decision threshold boundaries with correct and incorrect responses, and switching whether the starting point z on each trial was closer to the correct or incorrect boundary depending based on the rate of the presented stimulus (Wiecki *et al.*, 2016). This manipulation and our stimulus encoding scheme meant that values of z greater than 0.5 in a particular congruence condition indicated a systematic bias toward responding "fast," and values of z less than 0.5 indicated a systematic bias toward responding "slow," rather than a bias toward "correct" or "incorrect."

The joint posterior distribution of each model parameter was obtained via MCMC sampling. Four independent chains were run for 20 000 iterations each. The distributions of the final 10 000 iterations of each of the four chains (assumed to be represent steady-state probabilities) were combined, resulting in 40 000 posterior samples per parameter, per model. Marginal posterior distributions associated with each parameter within a model were characterized by examination of the 95% highest-density region (HDR) (Kruschke, 2013), which were obtained in R using the "hdrcde" package, version 3.1 (Hyndman, 2015).

3. Logistic regression modeling of task performance

A set of mixed-effect logistic regression models were fitted to the response data to analyze effects of the experimental manipulations on accuracy. This analysis was performed using the "lme4" package in R (Bates *et al.*, 2015). Separate sets of models were fitted to the auditory and visual tasks. In each model, the dependent variable was the binary response outcome (correct or incorrect), and the fixed effects included the various combinations of the experimental manipulations (congruence, modulation rate of the stimulus in the attended modality, and the interaction between these two factors). A random effect of participant (i.e., an "intercept-only" random effect) was included in each model. The best-fitting model was chosen from the set of candidate models on the basis of chi-squared tests reported by running the "anova" function on the lme4 model objects in R. Differences in levels of fixed effects in the best-fitting model were quantified using odds ratios. Odds ratios and their associated confidence intervals were computed and compared using generalized linear hypotheses tests, as implemented in R using the "glht" function from the "multcomp" package (Hothorn *et al.*, 2016).

III. RESULTS

A. Accuracy data

Proportion correct scores are shown in Fig. 3 for each task (visual task and auditory task in left and right panels, respectively), for trials with no opposite modality modulation, congruent modulation, and incongruent modulation. Performance was better overall in the auditory task than the visual task (compare distributions of points in the right panel to those in the left panels). In both tasks, mean proportion correct was lowest on incongruent trials (mean performance, visual task: 59.34%; mean performance, auditory task: 74.49%), highest on congruent trials (visual task: 75.18%; auditory task: 83.33%), and intermediate on unmodulated trials (visual task: 67.08%; auditory task: 79.79%).

B. DDM results

1. Reaction times

We sought to confirm suitability of our RT data to fitting with a "standard" DDM (in which all parameters are assumed to be time-invariant over the course of a trial), rather than using a more complicated model with time-varying parameters (e.g., see Milosavljevic *et al.*, 2010; White *et al.*, 2011). Briefly, a model with time-varying drift rates may be necessary to account for "conflict" data in which error RTs are faster than correct RTs on incongruent trials (White *et al.*, 2011). To investigate this possibility, RTs were pooled over all subjects and then grouped to form separate RT distributions for each task, modulation rate, congruence condition, and correct/incorrect responses. Results are shown in Fig. 4.

Considering incongruent trials only (Fig. 4, bottom panels), there is a slight tendency for shorter RTs for errors when participants were presented with fast visual stimuli in the visual task (Fig. 4, lower half of bottom left panel), and for slow auditory stimuli in the auditory task (Fig. 4, upper half of the bottom right panel). However, errors tend to have longer RTs for slow visual stimuli and fast auditory stimuli (Fig. 5, upper half of the bottom left panel and

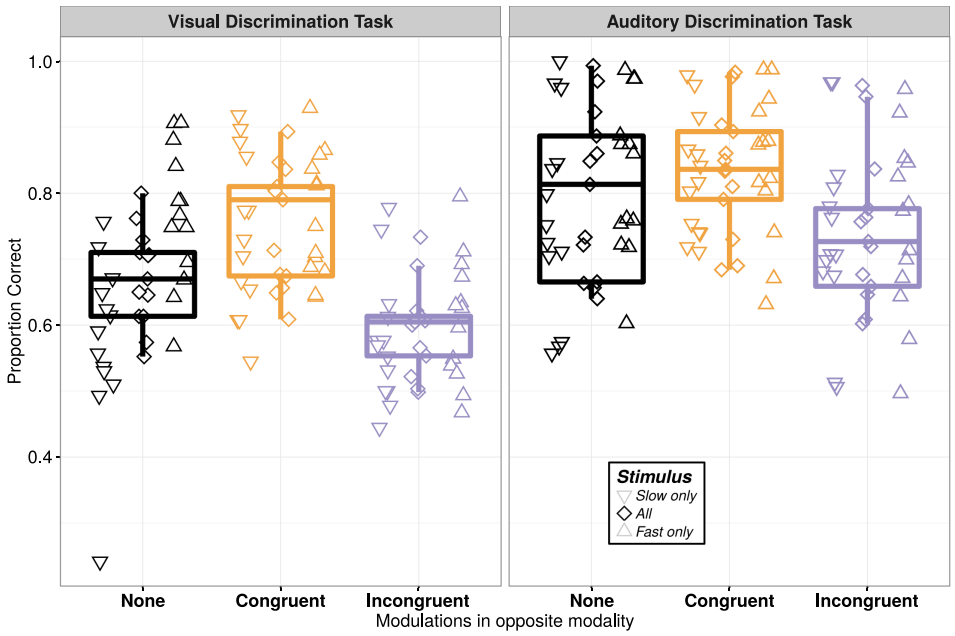

FIG. 3. (Color online) Proportion correct data for the visual discrimination task (left panel) and the auditory task (right panel). Each data point is from a single participant. Points are broken down based on discriminated stimulus rate (slow and fast, downward and upward facing triangles, respectively), and combined across the two rates (diamonds). Box plots illustrate median, 25th percentile (bottom), 75th percentile (top), and the lowest/highest values that are within 1.5× the interquartile (0.25–0.75) range (top and bottom whiskers).

lower half of the bottom right panel, respectively). Additionally, shorter error RTs are not confined to incongruent trials in the visual task; a tendency for fast errors is seen for fast visual stimuli in the visual task, regardless of what type of auditory stimulus is presented (see Fig. 4, bottom half of each of the left panels).

Overall, we concluded that while additional insights into the data might be possible using a more sophisticated decision-making model, there were no systematic, reliable changes in reaction times that would invalidate the interpretation of a standard seven-parameter DDM without time-varying parameters.

2. DDM parameter estimates

Effects of experimental manipulations on the posterior distributions of the parameters of interest (v, a, z, and t) are shown in Fig. 5. Since the unmodulated condition was used as a reference condition, the parameter values associated with the congruent and incongruent conditions were expressed as a change relative to that condition; to aid visual comparison across conditions, the congruent and incongruent parameter distributions reported below and shown in each panel of Fig. 5 were shifted by the mean value of the posterior distribution of the unmodulated condition for that task. Therefore, congruent and incongruent distributions can each be considered significantly different from the unmodulated condition if the 95% HDR of the congruent/incongruent posterior distribution does not include the mean value of the unmodulated distribution (indicated by the blue dashed line in each panel). The congruent and incongruent distributions can be considered different from one another if the 95% HDRs of the distributions do not overlap.

a. Drift rate (v). Results for v (Fig. 6; upper left panel) mimic the overall pattern of results evident from the raw accuracy data (Fig. 3). In both tasks, mean posterior v was smallest in the incongruent condition, intermediate in the unmodulated condition, and largest in the congruent condition. The congruent and incongruent 95% HDRs do not overlap with each other and do not include the mean posterior value of v in the unmodulated condition, meaning that in all congruence conditions, the values of v were all credibly different from one another.

b. Non-decision time (t). Non-decision time distributions (Fig. 5, upper right panel) indicated that the presence of an opposite-modality stimulus modulation led to shorter non-decision times in both tasks (congruent and incongruent distributions are shifted to the left of the dashed blue line). We note that although the tail end of the 95% HDR for the congruent condition in the auditory task included the mean of the unmodulated condition, it is reasonable to interpret the congruent change in non-decision time relative to the unmodulated condition as nonzero and negative given the values of the unmodulated mean and the upper end of the congruent 95% HDR (both numbers round to approximately 0.813). Non-decision times were also slightly shorter in the auditory task when visual stimuli were incongruent, compared to when the visual stimuli were congruent (incongruent distributions shifted to the left of the congruent distribution in the bottom plot only).

c. Decision thresholds (a). Decision thresholds (Fig. 5, lower left panel) were similar across the three conditions within either task; the exception was a shift toward a higher decision threshold in the auditory task for incongruent stimuli only.

d. Bias (z). Model fits of the bias parameter z (Fig. 5, bottom right) indicated that there was no significant bias when opposite modality modulations were present, in either task (bottom right panel; 95% HDR includes 0.5 for congruent and incongruent distributions for each task). However,

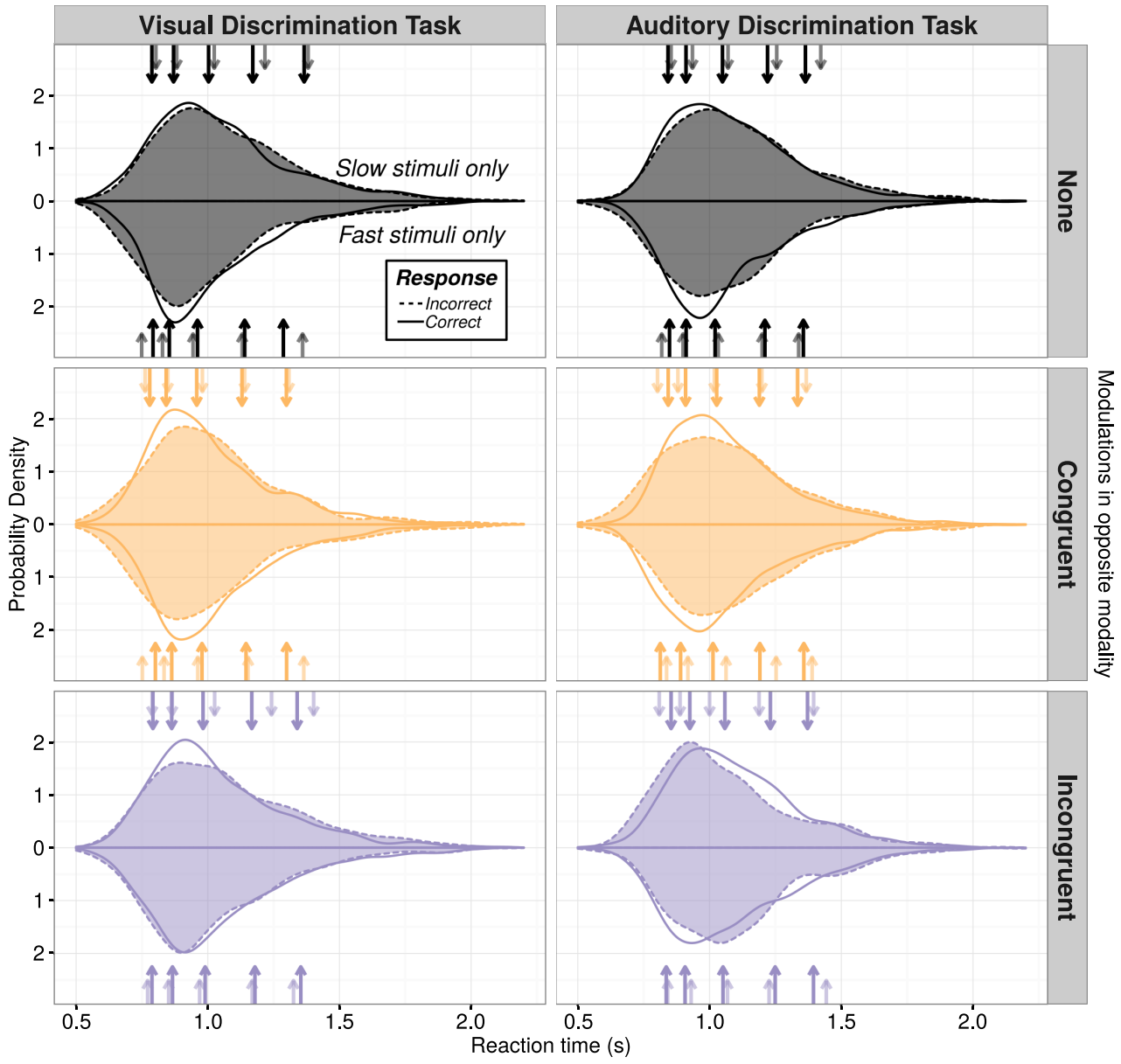

FIG. 4. (Color online) Reaction time distributions in the visual discrimination task (left panels) and auditory discrimination task (right panels), combined across all subjects and responses. Data are broken down by the opposite-modality modulations (top panels: unmodulated, middle panels: congruent, bottom panels: incongruent). Distributions are also broken down by stimulus rate and response: the top distributions in each panel show the reaction time probability densities for correct/incorrect (solid/dashed) responses to slow stimuli, and the lower, inverted distributions in each panel are the reaction time probability densities for correct/incorrect (solid/dashed) responses to fast stimuli. Arrows indicate the 12.5, 25, 50, 75, and 87.5 percentiles for each distribution, with dark and light arrows indicating the quantiles for correct and incorrect responses, respectively.

examination of the unmodulated results for each task indicates that there was a bias toward responding "fast" when the opposite-modality stimulus was unmodulated (lower right panel; unmodulated distributions are shifted to the right of 0.5).

3. Comparisons across tasks

Within each task, the differences between each congruence condition were subtracted from one another and then divided by the pooled standard deviation estimate to derive posterior distributions of an effect size similar to Cohen's *d* (Cohen, 1992). Results are shown in Fig. 6. From the plots, it becomes clear that there is an overall benefit of across-modality congruence in the modulation rates relative to when the unattended stimulus is unmodulated (Fig. 6, left panel; neither distribution includes 0), and that this benefit is larger for the visual task than it is for the auditory task (Fig. 6, left panel; 95% HDRs do not overlap). Comparing when the task-irrelevant sensory input is unmodulated and when it is incongruent, effect sizes are nonzero but similar across tasks (Fig. 6, middle panel; neither distribution includes 0, but 95% HDRs are overlapping). Finally, examination of the effect size comparing congruence vs incongruence for each task indicates that the HDRs are slightly overlapping. However, the HDRs suggest that the congruence vs incongruence effect is larger for the visual task than for the auditory task (Fig. 6, right panel; neither distribution includes 0, 95% HDRs exhibit some overlap).

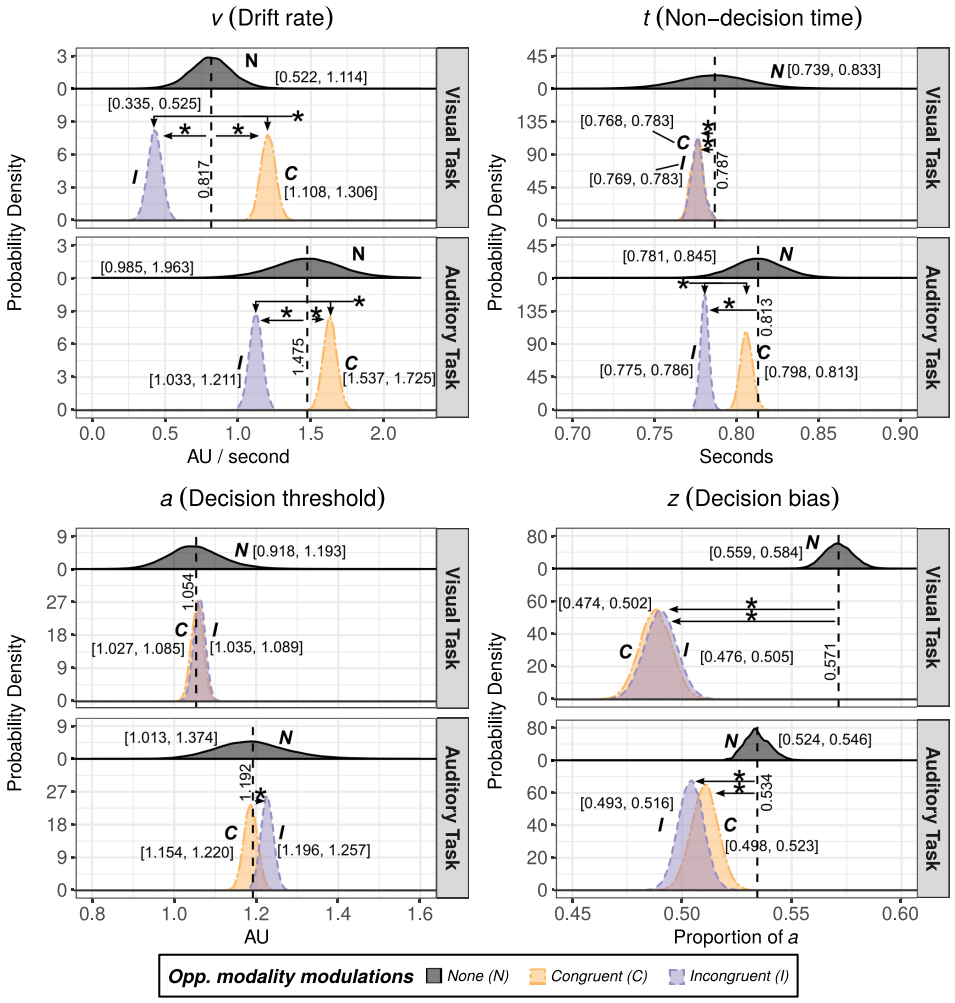

FIG. 5. (Color online) Posterior distributions of DDM parameters permitted to vary as a function of congruence. Parameters derived from data on the visual discrimination task occupy the top half of each panel, and parameters from the auditory task occupy the bottom half of each panel. Within each plot, the distribution near the top of the plot is the reference (unmodulated) condition. The mean of the unmodulated distribution is indicated in each panel by the vertical dashed line. The 95% HDR of each distribution is indicated in square brackets next to the appropriate curve. Arrows with asterisks indicate that there is no overlap in the 95% highest density regions for the two distributions being compared (when comparing congruent and incongruent conditions to one another), or that there is no overlap between a highest density region and the mean of the reference condition (when comparing congruent or incongruent distributions to the unmodulated condition); see text for details. AU = Arbitrary Units.

C. Logistic regression analysis of task performance

1. Logistic regression on accuracy, visual task

For the visual task, the logistic regression model that included covariates of stimulus rate, audiovisual congruence, and the interaction fit the response data best (Table I). Comparisons of odds ratios from this model (Table II) indicated that, regardless of the stimulus rate, odds of a correct response when participants identified visual stimuli were 2.1 times greater when the auditory modulations were congruent with the visual modulations compared to when they were incongruent (fast/congruent > fast/incongruent; slow/congruent > slow/incongruent). The patterns of odds ratios were different across stimulus rates in the unmodulated conditions. For fast stimuli, odds of a correct response were 2.1 times greater when the auditory stimuli were unmodulated compared to when the auditory stimuli were incongruently modulated (fast/unmodulated > fast/incongruent). However, there was no statistically significant difference in response accuracy when visual stimuli were fast and paired

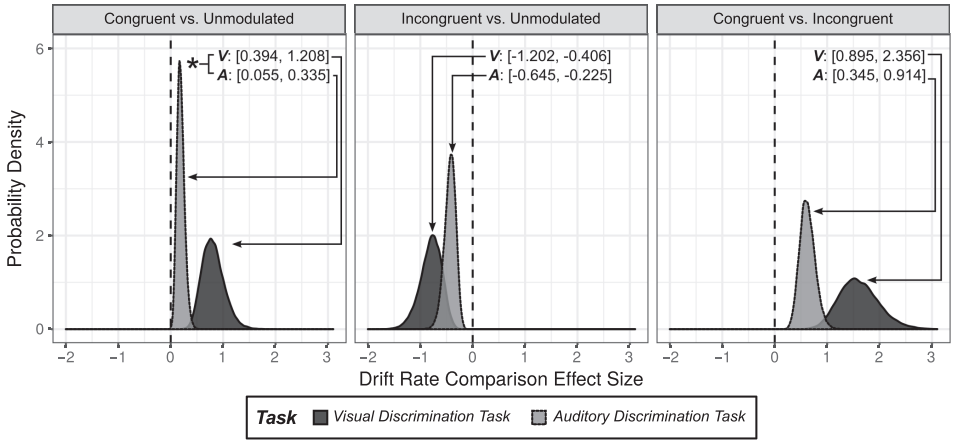

FIG. 6. Effect sizes of congruence on drift rate, computed as the difference between two posterior distributions (from amongst the congruent/incongruent/unmodulated distributions shown in the upper left panel of Fig. 5), divided by the estimate of the pooled standard deviation of that parameter estimate (see Sec. II). The comparison of interest is labeled at the top of each panel; the numbers in brackets in each panel indicate the 95% HDR for the visual (V) and auditory (A) effect size distributions for each comparison.

TABLE I. Logistic regression model specification and selection for the visual discrimination task. Model specifications in Wilkinson/lme4 notation, with random effect of subject specified (i.e., separate intercept for each subject). Best fitting model (lowest AIC, significant result from Chi Squared test) is indicated in bold.

Model Specification	DF[a]	AIC[b]	Chi Squared[c]	ΔDF[d]	p value[e]
response ~ (1 | subj)	2	14538			
response ~ rate + (1 | subj)	3	14460	80.268	1	$\ll 0.001$[f]
response ~ congruence + (1 | subj)	4	14321	141.101	1	$\ll 0.001$[f]
response ~ congruence + rate + (1 | subj)	5	14230	93.452	1	$\ll 0.001$[f]
response ~ congruence + rate + congruence:rate + (1 | subj)	7	14158	75.304	2	$\ll 0.001$[f]

[a]Degrees of freedom (DF).
[b]Akaike Information Criterion (AIC).
[c]Chi Squared value from testing against the previously listed model.
[d]Change in degrees of freedom vs. previously listed model (ΔDF).
[e]p value from Chi Squared test.
[f]Significant p values.

with auditory stimuli that were either congruent or unmodulated (i.e., fast/congruent ≈ fast/unmodulated; odds ratio between conditions of approximately 1, p approximately 1). For slow visual modulations, odds of a correct response were 2.1 times greater when these were paired with congruent auditory modulations compared to when there were no auditory modulations (slow/congruent > slow/unmodulated). There was no statistically significant difference when comparing performance for judgments of slow auditory modulations when the visual modulations were either not present or incongruent (slow/unmodulated ≈ slow/incongruent; odds ratio estimate to be approximately 1; p approximately 1).

2. Logistic regression on accuracy, auditory task

For the auditory task, a model containing effects of congruence, rate, and the interaction between the two factors best explained the data (Table III). The general trends from this model are similar to those observed in the visual task (Table IV).

Odds of a correct response were greater when the accompanying visual modulations were congruent with the auditory modulations compared to when they were incongruent; odds ratios were approximately 1.9 and 1.7 when participants judged fast and slow auditory modulations, respectively. Odds of a correct response when judging auditory stimuli were 1.6 times greater when fast auditory modulations were paired with unmodulated visual stimuli compared to when fast auditory modulations were paired with incongruent visual modulations (fast/unmodulated > fast/incongruent). However, they were equally likely to respond correctly when fast auditory stimuli were paired with either congruent or unmodulated visual modulations (fast/congruent ≈ fast/unmodulated; odds ratio approximately 1.1, $p = 0.528$, 95% confidence interval of the odds ratio includes 1). For the slow auditory stimuli, participants were about 1.4 times as likely to respond correctly when the audio and visual modulations were congruent compared to when there were no visual modulations (slow/congruent > slow/unmodulated), but were about equally likely respond correctly when there were no visual modulations was unmodulated compared to when the visual modulations were incongruent (slow/unmodulated ≈ slow/incongruent; odds ratio approximately 1.2; $p = 0.102$, 95% confidence interval of the odds ratio includes 1).

3. Comparisons across tasks

We considered whether an individual's performance on the visual task was related to their performance on the auditory task (Fig. 7). Simple linear regressions indicated that regardless of the stimulus rate or cross-modal condition,

TABLE II. Summary of logistic regression fixed effects when compared using generalized linear hypothesis tests. Data are compared separately for fast and slow stimuli due to the significance of the interaction term in the model selection process.

Stimulus rate	Modulation comparison	Estimate[a]	SE[b]	z value[c]	p value[d]	OR [95% CI][e]
Fast	Congruent vs Incongruent	0.740	0.072	10.234	<0.001[f]	2.096 [1.737, 2.529]
	Unmodulated vs Incongruent	0.745	0.071	10.509	<0.001[f]	2.106 [1.752, 2.532]
	Congruent vs Unmodulated	−0.005	0.078	−0.063	~1	0.995 [0.814, 1.217
Slow	Congruent vs Incongruent	0.747	0.069	10.745	<0.001[f]	2.110 [1.761, 2.527]
	Unmodulated vs Incongruent	−0.001	0.067	−0.010	~1	0.999 [0.841, 1.188]
	Congruent vs Unmodulated	0.747	0.069	10.851	<0.001[f]	2.111 [1.765, 2.525]

[a]Estimate is the coefficient corresponding to the comparison in the previous column.
[b]Standard error of the estimate (SE).
[c]z Value is the z value for the modulation comparison listed.
[d]p Value is the p value adjusted for multiple comparisons via single-step method.
[e]Odds ratio (OR); 95% confidence interval for the odds ratio (95% CI).
[f]Significant p values.

TABLE III. Logistic regression model specification and selection for the auditory discrimination task. Model specifications in Wilkinson/lme4 notation, with random effect of subject specified (i.e., separate intercepts for each subject). Best fitting model (lowest AIC, significant result from Chi Squared test) is indicated in bold.

Model specification	DF[a]	AIC[b]	Chi Squared[c]	ΔDF[d]	*p* value[e]
response ~ (1 | subj)	2	11059			
response ~ rate + (1 | subj)	3	11041	19.951	1	<0.001[f]
response ~ congruence + (1 | subj)	4	10961	81.428	1	<0.001[f]
response ~ congruence + rate + (1 | subj)	5	10947	16.711	1	<0.001[f]
response ~ congruence + rate + congruence:rate + (1 | subj)	7	10943	7.864	2	0.020[f]

[a]Degrees of freedom (DF).
[b]Akaike Information Criterion (AIC).
[c]Chi Squared value from testing against the previously listed model.
[d]Change in degrees of freedom vs. previously listed model (ΔDF).
[e]*p* value from Chi Squared test.
[f]Significant *p* values.

proportions correct on the tasks were not correlated with one another.

Finally, we considered a logistic regression model to examine the accuracy data combined across tasks. Task (visual or auditory), condition (opposite modality was congruent, incongruent, or unmodulated), and task-relevant stimulus rate (fast or slow), as well as all combinations of interactions between these factors, were included as fixed effects in this model. A random effect of task-within-participant was included in the model fit, i.e., the effect of task (visual task and auditory task) was clustered within participant.

Tests of the fixed effect model parameters (Table V) confirmed that performance was generally better in the auditory task than on the visual task; participants were about 2.3 times more likely to respond correctly in the auditory task compared to the visual task. Combining across tasks, participants were 1.9 times more likely to respond correctly when stimuli were congruent compared to when they were incongruent when data are considered independently of task. The rate-dependent effects indicated in the models fit separately for each task also held for the combined model: there was no difference between performance when the task-irrelevant stimulus was incongruent or unmodulated when discriminating slow stimuli (odds ratio approximately 1.1, 95% confidence interval includes the value one), or between performance when the task irrelevant stimulus was congruent and when the task irrelevant stimulus was unmodulated when discriminating fast stimuli (odds ratio approximately 1.1, 95% confidence interval includes the value one).

Odds ratios obtained from logistic regression the odds ratios may be interpreted directly as effect sizes (Fleiss *et al.*, 1994). When performing an effect size comparison on the logistic regression models that was analogous to the effect size comparison performed for the drift rate parameter in the DDM, the larger effects of congruence in the visual task relative to the auditory task are absent: the 95% confidence intervals of the odds ratios for the comparisons involving congruence overlap when they are compared across tasks (compare OR and 95% CI in Table II and Table IV).

IV. DISCUSSION

A. Congruence effects in both tasks suggest audiovisual effects of modulation rate are bi-directional, but asymmetric

Congruency effects were observed in both the visual task and the auditory task: judgment accuracy was highest when audio and visual stimuli were congruent, lowest when they were incongruent, and intermediate when the task-

TABLE IV. Summary of logistic regression fixed effects when compared using generalized linear hypothesis tests. Data are compared separately for fast and slow stimuli due to the significance of the interaction term in the model selection process.

Stimulus rate	Modulation comparison	Estimate[a]	SE[b]	*z* value[c]	*p* value[d]	OR [95% CI][e]
Fast	Congruent vs Incongruent	0.617	0.084	7.352	<0.001[f]	1.853 [1.490, 2.304]
	Unmodulated vs Incongruent	0.489	0.083	5.881	<0.001[f]	1.631 [1.314, 2.024]
	Congruent vs Unmodulated	0.127	0.088	1.454	0.528	1.136 [0.905, 1.427]
Slow	Congruent vs Incongruent	0.539	0.082	6.557	<0.001[f]	1.715 [1.385, 2.124]
	Unmodulated vs Incongruent	0.179	0.077	2.324	0.102	1.196 [0.979, 1.462]
	Congruent vs Unmodulated	0.360	0.085	4.263	<0.001[f]	1.434 [1.151, 1.786]

[a]Estimate is the coefficient corresponding to the comparison in the previous column.
[b]Standard error of the estimate (SE).
[c]*z* Value is the *z* value for the modulation comparison listed.
[d]*p* Value is the *p* value adjusted for multiple comparisons via single-step method.
[e]Odds ratio (OR); 95% confidence interval for the odds ratio (95% CI).
[f]Significant *p* values.

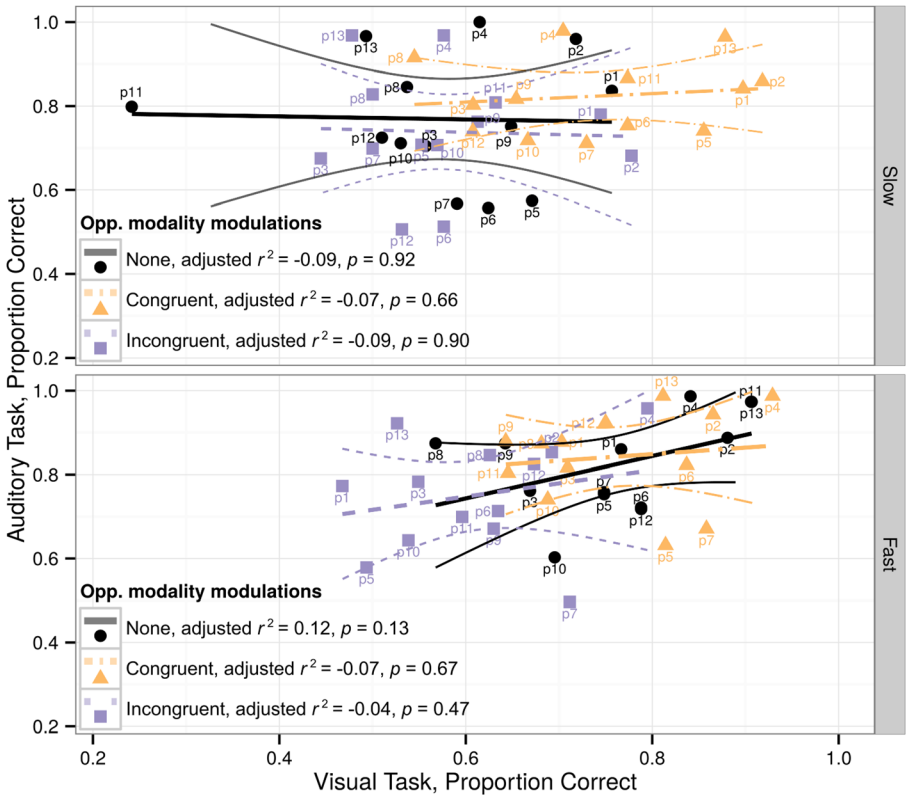

FIG. 7. (Color online) Correlations between proportion correct for the auditory (y axis) and visual (x axis), tasks, plotted separately for the two stimulus rates (slow and fast; labels at right). Text labels indicate participant identities to facilitate comparisons of the same participant across plots. Shaded areas are 95% confidence intervals around the regression lines. Adjusted r^2 and p values (from an F-test) for each condition are reported in each figure legend.

irrelevant modality was unmodulated, regardless of which sensory input (auditory or visual) the participants were judging. DDM drift rates paralleled the accuracy findings, indicating that the quality of evidence contributing to a decision about modulation rate was highest on congruent trials and lowest on incongruent trials. Our observation of congruency effects in both tasks suggests that audiovisual interactions for temporal modulations occur bi-directionally.

Bi-directional audiovisual interactions may arise on congruent trials via neural mechanisms that act to favor perceptual binding of sensory inputs with similar temporal properties (Bizley *et al.*, 2016). Binding of information with similar spatiotemporal properties and the formation of visual "objects" or auditory "streams" is the basis of scene organization within a single modality (Treisman and Gelade, 1980; Luck and Vogel, 1997; Shinn-Cunningham, 2008). Audiovisual object formation appears to work in similar ways (Bizley *et al.*, 2012; Bizley *et al.*, 2016). As such, temporal modulations in different modalities at the same rate may force the percept of a bound object in a stimulus-driven manner (Koelewijn *et al.*, 2010; Talsma *et al.*, 2010). However, we observed cross-modal interactions when the task-irrelevant input was incongruent, not only when it was congruent. Specifically, drift rates were slower for incongruent trials than when the irrelevant input was unmodulated (see Fig. 7); the interactions arise even when the temporal

TABLE V. Summary of accuracy logistic regression fixed effects when comparisons of interest are tested using generalized linear hypothesis tests. In Wilkinson/lme4 notation, the model was specified as response ~ condition + stimulus + task + condition:stimulus + condition:task + task:stimulus + condition:task:stimulus + (1 + task | subj). Parameters listed were obtained by combining over interactions when performing the hypothesis tests. Abbreviations as in Table II.

Comparison	Estimate[a]	SE[b]	z value[c]	p value[d]	OR [95% CI][e]
Visual vs Auditory	−0.818	0.249	−3.288	0.006[f]	0.441 [0.232, 0.840]
Congruent vs Incongruent	0.660	0.039	17.105	<0.001[f]	1.936 [1.752, 2.140]
Slow/Congruent vs Slow/Unmodulated	0.554	0.055	10.160	<0.001[f]	1.740 [1.511, 2.003]
Slow/Unmodulated vs Slow/Incongruent	0.089	0.051	1.753	0.337	1.093 [0.958, 1.247]
Fast/Congruent vs Fast/Unmodulated	0.061	0.058	1.049	0.808	1.063 [0.914, 1.237]
Fast/Unmodulated vs Fast/Incongruent	0.600	0.145	4.140	<0.001[f]	1.822 [1.251, 2.651]

[a]Estimate is the coefficient corresponding to the comparison in the previous column.
[b]Standard error of the estimate (SE).
[c]z Value is the z value for the modulation comparison listed.
[d]p Value is the p value adjusted for multiple comparisons via single-step method.
[e]Odds ratio (OR); 95% confidence interval for the odds ratio (95% CI).
[f]Significant p values.

fluctuations in the two modalities do not match. This may be because temporal coherence in modulations across modalities may only affect binding for stimulus modulation rates below about 4 Hz or so (Fujisaki and Nishida, 2005). We used rates of 6 and 7 Hz, which may be too rapid to influence binding directly; instead, it could be that the auditory and visual inputs are bound even in incongruent modulation trials because the auditory and visual inputs turn on and off together. Alternatively, interactions between auditory and visual information may occur at later processing stages that are more related to audiovisual congruence/incongruence resolution (e.g., Hein *et al.*, 2007; Noppeney *et al.*, 2010), and completely unrelated to cross-modal binding. Mistakes on incongruent trials may reflect cognitive mechanisms. In this sense, our task and results may have more in common with within-modality conflict resolution tasks, such as Stroop Tasks (e.g., Vendrell *et al.*, 1995; Leung *et al.*, 2000) or Eriksen Flanker Tasks (e.g., van Veen and Carter, 2002). Indeed, one potential explanation for the pattern of drift rate results, in which higher drift rates were found for congruent stimuli and lower drift rates were found for incongruent stimuli compared to the unmodulated condition, is that the observed drift rates in the fitted DDM correspond to differently weighted drift processes from each modality, with the weighting determined by the focus of modality-specific attention. In this view, facilitation or interference effects are likely to be cognitive rather than perceptual. More generally speaking, however, distinguishing true perceptual binding effects and from cognitive conflict resolution mechanisms will require different paradigms and stimuli (e.g., see Bizley *et al.*, 2016) or may require disambiguation using functional imaging methods (see Sec. IV B, below).

Despite observations of bi-directionality, and independent of the discussion about the mechanisms by which it arises, the DDM drift rate effect sizes indicate that the increase in evidence quality (drift rate) that comes about due to congruent modulations tended to be larger for the visual task than in the auditory task (Fig. 7). The asymmetry in the level of temporal congruence benefit is consistent with findings demonstrating that the auditory system is more suited to temporal processing than the visual system (Welch and Warren, 1980; Recanzone, 2002; Michalka *et al.*, 2015). In this view, the percept of time-varying features in a scene should be dominated by information encoded in the sensory system optimized for temporal information (i.e., the auditory system). In contrast, inputs to the visual system may provide a more reliable source of information for spatial judgments; for example, visual stimulus motion has been shown to affect judgments regarding direction of auditory apparent motion (Soto-Faraco *et al.*, 2002; Soto-Faraco *et al.*, 2004). These rules, however, may not hold for inputs that are degraded in some way; for instance, while the visual system is suited for spatial judgments, spatial information will be extracted from auditory information if a visual input provides sufficiently ambiguous information regarding position (Alais and Burr, 2004).

Changes in each task were also observed in the non-decision time parameter (congruent and incongruent non-decision times were faster than those on unmodulated trials) and the bias parameter (bias values were shifted toward responding "fast" more often on the unmodulated condition in each task). Non-decision times are conceptualized as including stimulus encoding times (Ratcliff and McKoon, 2008), and thus decreases in this parameter when modulations were present may be a behavioral consequence of shorter latencies arising from the engagement of neural populations sensitive to multi-modal temporal modulations (e.g., Meredith *et al.*, 1987). More investigation is needed, however, since non-decision times also comprise motor response times and possibly sources of variance in RT distributions (Ratcliff and McKoon, 2008).

It is harder to speculate on what caused the changes in bias observed in both tasks. For the visual discrimination task, the change in bias due in the unmodulated condition bias could be partly due to a single subject responding "fast" a disproportionate number of times when a slow visual stimulus was presented with an unmodulated auditory stimulus (see the leftmost set of points in the left panel of Fig. 3). For the auditory discrimination task, the bias might be explained by participants doing slightly worse overall when presented with slow auditory stimuli and unmodulated visual stimuli (mean accuracy: 76.91%), compared to when they were presented with fast auditory stimuli and unmodulated visual stimuli (mean accuracy: 82.66%). Here, we note that if computing the bias using DDM was not an option (perhaps due to a lack of RT data or a lack of computational power), the bias in participant responses could have been deduced from an analysis using traditional signal detection theory rather than performing a logistic regression on correct and incorrect responses. Alternatively, a systematic bias could have been determined by fitting logistic regression models with the same fixed effects, but with the participant response (fast/slow) as the dependent variable in the models.

B. Limitations and future work

A limitation in drawing comparisons between findings on an auditory task and a visual task is that the amount of information conveyed via each sense and the strategies employed by participants in utilizing these two sources of sensory information may differ from one another. Although identical physical stimuli were employed during both the auditory and visual tasks in the current study, it is likely that the amount of visual information perceived during the auditory task was less than the amount perceived during the visual task. A strategy that participants could have employed on auditory trials was to fix their gaze on a portion of the screen outside the game window or on some portion of the game window that would prevent the image of the moving fish entering the fovea. While an infrared camera was utilized to ensure that participants were not closing their eyes, turning their heads, or otherwise blatantly disregarding instructions, controlling for gaze position could only have been by tracking eye movements. Given this problem, it is difficult to argue that the effects of visual inputs on auditory modulation discrimination are fundamentally weaker than the effects of auditory inputs on visual modulation discrimination based on the current results alone. Furthermore, the

differences in strategy and stimulus information perceived (or utilized) by players may have contributed to the lack of correlation between performance levels on each task. On the other hand, the fact that congruence and incongruence effects were observed in the auditory task lends credence to the assertion that the audiovisual interactions arising from temporal modulations in each sensory modality are automatic.

Proportion correct data and results of the logistic regression analysis of accuracy suggest that there may be interactions between audiovisual interactions and stimulus rates. Specifically, performance was no better when the auditory and visual stimuli were both at 7 Hz (i.e., congruent) than when the task-irrelevant stimuli were unmodulated. In contrast, judgments of slow stimuli in both tasks were likely to benefit from the presence of congruent, task-irrelevant modulations, but performance on incongruent and opposite-modality unmodulated trials was similar. The simplest explanation for this finding in the opposite-modality-unmodulated case is there is more information available to the observer in a fixed amount of time for a fast modulation rate than for a slow modulation rate. One (admittedly speculative) possibility is that the benefit of congruence at 6 Hz is due to the fusion of audio and visual stimuli enhancing perception when the modulations are congruent, but the effects of incongruence observed for 7 Hz stimuli are cognitive-level confusion effects that interfere with information accumulation. That effects differ at these two rates may not be surprising; for instance, previous studies of audiovisual interactions in speech have suggested that perceptual binding of audio and visual stimuli may be weak for temporal modulation rates above approximately 7 Hz (Chandrasekaran *et al.*, 2009). The 6–7 Hz rates used in the present experiment may straddle some critical rate below which task performance is dictated by perceptual-level binding, and above which performance is dictated by other mechanisms. We note that our findings hinting at rate-specific effects are at odds with the previous *Fish Police!* experiments, in which no rate effects were found. The differences between the previous results and the current set of results may be due to some combination of the different stimulus rates used in each experiment (6 and 7 Hz in the current study, vs 6 and 8 Hz in the previous studies) and the different trial blocking employed in each experiment (congruent, incongruent, and unmodulated trials were intermingled, in random order, in the current experiment, compared to having the three conditions in separate blocks of trials in Sun *et al.*, 2016). Taken together, these results point to a need for additional experiments to examine interactions between cross-modal temporal modulations and the rates at which they occur. We obtained sensible and easily interpretable results using a "standard" seven-parameter DDM. Still, alternative formulations of the DDM may provide further insights into the strategies participants use when they are successfully able to resolve incongruence across auditory and visual modalities. For example, versions of the DDM have been developed to explicitly deal with "conflict" tasks (White *et al.*, 2011; Ulrich *et al.*, 2015). It is also possible to model time pressures, such as those imposed on each trial within the game in the present study, explicitly within the DDM framework. This can be done by making decision boundaries dependent on time, and by collapsing the distance between the two boundaries as time increases (Milosavljevic *et al.*, 2010). These more complex models can account for time-varying changes in evidence available to the participant, or multiple-stage decisional models in which a decision is made once some subset of the available evidence has been selected. The downside to fitting time-varying DDMs is that the readily available software packages for fitting standard DDMs (e.g., HDDM) must be modified extensively to be adapted for that purpose. Furthermore, more complex models may not offer additional insights into the data that are not available with a simpler model; for example, DDM models in which boundaries are allowed to collapse do not always fit the data better than a "simple" DDM without time varying parameters (Milosavljevic *et al.*, 2010).

Finally, although sequential sampling models provide additional insights into decision making relative to independent analyses of performance and RT, the models cannot directly identify changes in decision-making processes that occur at the perceptual level vs those that are more cognitive in nature. In other words, fitting model parameters to behavioral data alone cannot distinguish between cross-modal interactions that occur at sites associated with sensory-level cross-modal processing (e.g., in superior colliculus; Meredith and Stein, 1986; Meredith *et al.*, 1987) or those that occur in brain areas more directly associated with cognition and decision making (e.g., prefrontal cortex; Euston *et al.*, 2012). When using DDM or similar models, such distinctions can be drawn in humans with the aid of functional neuroimaging techniques such as fMRI or MEG/EEG. Sequential sampling models and their central theme of modeling information accumulation can be easily related to work seeking to identify candidates of information accumulation sites for audiovisual tasks in the brain (Noppeney *et al.*, 2010). Additionally, some recent studies have utilized regression methods to relate parameters from sequential sampling models to EEG (Cavanagh *et al.*, 2011) and local field potential (Herz *et al.*, 2016) data. Using similar techniques on behavioral and neural data from audiovisual tasks may be a particularly useful method for identifying the neural loci and time courses of various types of audiovisual interactions, including the temporal modulations utilized in the present experiment.

C. Conclusions

Audio-visual interactions for temporal modulations are bi-directional and obligatory; observers are affected by temporal fluctuations in a task-irrelevant sensory input even when they know that input will be uninformative. Although bi-directional, the influence of auditory information on visual judgments is larger than the influence of visual information on auditory judgments. When data were entered into a DDM, the main difference between congruent and incongruent auditory and visual temporal modulations emerges as differences in drift rate, which corresponds to quality of stimulus evidence available to the observer. Future studies should aim to resolve whether audiovisual interactions

involving temporal modulations arise because of neural mechanisms at early sensory integration sites, or in brain regions more closely associated with cognitive processing and information accumulation.

ACKNOWLEDGMENTS

This work was funded by CELEST, a National Science Foundation Science of Learning Center (SBE-0354378), and SL-CN: Engaging Learning Network, a National Science Foundation Collaborative Network (SMA/SBE-1540920). We would like to thank Lorraine Delhorne for conducting hearing screenings on the individuals who took part in this study. We would also like to thank Diego Fernandez-Duque and three anonymous reviewers for their comments on an earlier version of this manuscript.

[1]See supplementary material at http://dx.doi.org/10.1121/1.4979470 for supplementary figure, a copy of the source code of the game, and a video of gameplay.

Alais, D., and Burr, D. (**2004**). "The ventriloquist effect results from near-optimal bimodal integration," Curr. Biol. **14**, 257–262.
Bates, D., Mächler, M., Bolker, B., and Walker, S. (**2015**). "Fitting linear mixed-effects models using lme4," J. Stat. Softw. **67**, 1–48.
Bizley, J. K., Maddox, R. K., and Lee, A. K. C. (**2016**). "Defining auditory-visual objects: Behavioral tests and physiological mechanisms," Trends Neurosci. **39**, 74–85.
Bizley, J. K., Shinn-Cunningham, B. G., and Lee, A. K. C. (**2012**). "Nothing is irrelevant in a noisy world: Sensory illusions reveal obligatory within- and across-modality integration," J. Neurosci. **32**, 13402–13410.
Bogacz, R., Brown, E., Moehlis, J., Holmes, P., and Cohen, J. D. (**2006**). "The physics of optimal decision making: A formal analysis of models of performance in two-alternative forced-choice tasks," Psychol. Rev. **113**, 700–765.
Cavanagh, J. F., Wiecki, T. V., Cohen, M. X., Figueroa, C. M., Samanta, J., Sherman, S. J., and Frank, M. J. (**2011**). "Subthalamic nucleus stimulation reverses mediofrontal influence over decision threshold," Nat. Neurosci. **14**, 1462–1467.
Chandrasekaran, C., Trubanova, A., Stillittano, S., Caplier, A., and Ghazanfar, A. A. (**2009**). "The natural statistics of audiovisual speech," PLoS Computat. Biol. **5**(7), e1000436.
Cohen, J. (**1992**). "A power primer," Psychol. Bull. **112**, 115–159.
Denison, R. N., Driver, J., and Ruff, C. C. (**2013**). "Temporal structure and complexity affect audio-visual correspondence detection," Front. Psychol. **3**, 619.
Euston, D. R., Gruber, A. J., and McNaughton, B. L. (**2012**). "The role of medial prefrontal cortex in memory and decision making," Neuron **76**, 1057–1070.
Faraway, J. J. (**2014**). *Linear Models With R*, 2nd ed. (CRC Press, Boca Raton, FL).
Fleiss, J. L., Cooper, H., and Hedges, L. V., eds. (**1994**). *The Handbook of Research Synthesis* (Russell Sage Foundation, New York), pp. 245–260.
Forstmann, B. U., Ratcliff, R., and Wagenmakers, E.-J. (**2016**). "Sequential sampling models in cognitive neuroscience: Advantages, applications, and extensions," Annu. Rev. Psychol. **67**, 641–666.
Fujisaka, W., and Nishida, S. (**2005**). "Temporal frequency characteristics of synchrony-asynchrony discrimination of audio-visual signals," Exp. Brain Res. **166**(3–4), 455–464.
Gebhard, J. W., and Mowbray, G. H. (**1959**). "On discriminating the rate of visual flicker and auditory flutter," Am. J. Psychol. **72**, 521–529.
Goldberg, H., Sun, Y., Hickey, T. J., Shinn-Cunnigham, B., and Sekuler, R. (**2015**). "Policing fish at Boston's Museum of Science: Studying audiovisual interaction in the wild," i-Perception **6**(4), 1.
Green, D. M., and Swets, J. A. (**1966**). *Signal Detection Theory and Psychophysics* (Wiley, New York).
Hein, G., Doehrmann, O., Muller, N. G., Kaiser, J., Muckli, L., and Naumer, M. J. (**2007**). "Object familiarity and semantic congruency modulate responses in cortical audiovisual integration areas," J. Neurosci. **27**, 7881–7887.
Heitz, R. P. (**2014**). "The speed-accuracy tradeoff: History, physiology, methodology, and behavior," Front. Neurosci. **8**, 150.
Herz, D. M., Zavala, B. A., Bogacz, R., and Brown, P. (**2016**). "Neural correlates of decision thresholds in the human subthalamic nucleus," Curr. Biol. **26**, 916–920.
Hickey, T. J. (**2013**). fishgame, https://github.com/tjhickey724/fishgame (Last viewed January 4, 2017).
Hothorn, T., Bretz, F., Westfall, P., Heiberger, R. M., Schuetzenmeister, A., Scheibe, S., and Hothorn, M. T. (**2016**). Package "multcomp," http://cran.stat.sfu.ca/web/packages/multcomp/multcomp.pdf (Last viewed February 21, 2017).
Hyndman, R. J. (**2015**). Package "hdrcde," http://cran.stat.sfu.ca/web/packages/hdrcde/hdrcde.pdf (Last viewed February 21, 2017).
Koelewijn, T., Bronkhorst, A., and Theeuwes, J. (**2010**). "Attention and the multiple stages of multisensory integration: A review of audiovisual studies," Acta Psychol. (Amst.) **134**, 372–384.
Kruschke, J. K. (**2013**). "Bayesian estimation supersedes the t test," J. Exp. Psychol. Gen. **142**, 573–603.
Kubovy, M., and Yu, M. (**2012**). "Multistability, cross-modal binding and the additivity of conjoined grouping principles," Philos. Trans. R. Soc. Lond. B Biol. Sci. **367**, 954–964.
Leung, H.-C., Skudlarski, P., Gatenby, J. C., Peterson, B. S., and Gore, J. C. (**2000**). "An event-related functional MRI study of the stroop color word interference task," Cereb. Cortex **10**, 552–560.
Luck, S. J., and Vogel, E. K. (**1997**). "The capacity of visual working memory for features and conjunctions," Nature **390**, 279–281.
Maddox, R. K., Atilgan, H., Bizley, J. K., and Lee, A. K. (**2015**). "Auditory selective attention is enhanced by a task-irrelevant temporally coherent visual stimulus in human listeners," Elife **4**, e04995.
Marks, L. E. (**1987**). "On cross-modal similarity: Auditory-visual interactions in speeded discrimination," J. Exp. Psychol. Hum. Percept. Perform. **13**, 384–394.
Mathias, S. R. (**2016**). "Unified analysis of accuracy and reaction times via models of decision making," Proc. Mtgs. Acoust. **26**, 050001.
Matzke, D., and Wagenmakers, E.-J. (**2009**). "Psychological interpretation of the ex-Gaussian and shifted Wald parameters: A diffusion model analysis," Psychon. Bull. Rev. **16**, 798–817.
Meredith, M. A., Nemitz, J. W., and Stein, B. E. (**1987**). "Determinants of multisensory integration in superior colliculus neurons. I. Temporal factors," J. Neurosci. **7**, 3215–3229.
Meredith, M. A., and Stein, B. E. (**1986**). "Visual, auditory, and somatosensory convergence on cells in superior colliculus results in multisensory integration," J. Neurophysiol. **56**, 640–662.
Michalka, S. W., Kong, L., Rosen, M. L., Shinn-Cunningham, B. G., and Somers, D. C. (**2015**). "Short-term memory for space and time flexibly recruit complementary sensory-biased frontal lobe attention networks," Neuron **87**, 882–892.
Milosavljevic, M., Malmaud, J., Huth, A., Koch, C., and Rangel, A. (**2010**). "The drift diffusion model can account for the accuracy and reaction time of value-based choices under high and low time pressure," Judgm. Decis. Mak. **5**, 437–449.
Miranda, A. T., and Palmer, E. M. (**2013**). "Intrinsic motivation and attentional capture from gamelike features in a visual search task," Behav. Res. Methods **46**(1), 159–172.
Molholm, S., Martinez, A., Shpaner, M., and Foxe, J. J. (**2007**). "Object-based attention is multisensory: Co-activation of an object's representations in ignored sensory modalities," Eur. J. Neurosci. **26**, 499–509.
Noppeney, U., Ostwald, D., and Werner, S. (**2010**). "Perceptual decisions formed by accumulation of audiovisual evidence in prefrontal cortex," J. Neurosci. Off. J. Soc. Neurosci. **30**, 7434–7446.
Parise, C. V., Spence, C., and Ernst, M. O. (**2012**). "When correlation implies causation in multisensory integration," Curr. Biol. **22**, 46–49.
Patil, A., Huard, D., and Fonnesbeck, C. J. (**2010**). "PyMC: Bayesian stochastic modelling in Python," J. Stat. Softw. **35**, 1.
Ratcliff, R. (**1978**). "A theory of memory retrieval," Psychol. Rev. **85**, 59–108.
Ratcliff, R., and Childers, R. (**2015**). "Individual differences and fitting methods for the two-choice diffusion model of decision making," Decision **2**, 237–279.
Ratcliff, R., and McKoon, G. (**2008**). "The diffusion decision model: Theory and data for two-choice decision tasks," Neural Comput. **20**, 873–922.
Ratcliff, R., and Rouder, J. N. (**1998**). "Modeling response times for two-choice decisions," Psychol. Sci. **9**, 347–356.

Recanzone, G. H. (**2002**). "Auditory influences on visual temporal rate perception," J. Neurophysiol. **89**, 1078–1093.
Shams, L., Kamitani, Y., and Shimojo, S. (**2002**). "Visual illusion induced by sound," Cogn. Brain Res. **14**, 147–152.
Shinn-Cunningham, B. G. (**2008**). "Object-based auditory and visual attention," Trends Cogn. Sci. **12**, 182–186.
Shipley, T. (**1964**). "Auditory flutter-driving of visual flicker," Science **145**, 1328–1330.
Soto-Faraco, S., Lyons, J., Gazzaniga, M., Spence, C., and Kingstone, A. (**2002**). "The ventriloquist in motion: Illusory capture of dynamic information across sensory modalities," Cogn. Brain Res. **14**, 139–146.
Soto-Faraco, S., Spence, C., and Kingstone, A. (**2004**). "Cross-modal dynamic capture: Congruency effects in the perception of motion across sensory modalities," J. Exp. Psychol. Hum. Percept. Perform. **30**, 330–345.
Speckman, P. L., Rouder, J. N., Morey, R. D., and Pratte, M. S. (**2008**). "Delta plots and coherent distribution ordering," Am. Stat. **62**, 262–266.
Spence, C. (**2011**). "Crossmodal correspondences: A tutorial review," Atten. Percept. Psychophys. **73**, 971–995.
Spence, C., and Driver, J. (**1997**). "On measuring selective attention to an expected sensory modality," Percept. Psychophys. **59**, 389–403.
Spence, C., and Squire, S. (**2003**). "Multisensory integration: Maintaining the perception of synchrony," Curr. Biol. **13**, R519–R521.
Sun, Y., Shinn-Cunningham, B., Hickey, T. J., and Sekuler, R. (**2016**). "Catching audiovisual interactions with a first-person fisherman video game," Perception. in press.
Talsma, D., Senkowski, D., Soto-Faraco, S., and Woldorff, M. G. (**2010**). "The multifaceted interplay between attention and multisensory integration," Trends Cogn. Sci. **14**, 400–410.
Treisman, A. M., and Gelade, G. (**1980**). "A feature-integration theory of attention," Cognit. Psychol. **12**, 97–136.
Ulrich, R., Schröter, H., Leuthold, H., and Birngruber, T. (**2015**). "Automatic and controlled stimulus processing in conflict tasks: Superimposed diffusion processes and delta functions," Cogn. Psychol. **78**, 148–174.
van Veen, V., and Carter, C. S. (**2002**). "The anterior cingulate as a conflict monitor: fMRI and ERP studies," Physiol. Behav. **77**, 477–482.
Vendrell, P., Junqué, C., Pujol, J., Jurado, M. A., Molet, J., and Grafman, J. (**1995**). "The role of prefrontal regions in the Stroop task," Neuropsychologia **33**, 341–352.
Voss, A., Nagler, M., and Lerche, V. (**2013**). "Diffusion models in experimental psychology: A practical introduction," Exp. Psychol. **60**, 385–402.
Wagenmakers, E.-J. (**2009**). "Methodological and empirical developments for the Ratcliff diffusion model of response times and accuracy," Eur. J. Cogn. Psychol. **21**, 641–671.
Washburn, D. A. (**2003**). "The games psychologists play (and the data they provide)," Behav. Res. Methods, Instrum., Comput. **35**(2), 185–193.
Welch, R. B., and Warren, D. H. (**1980**). "Immediate perceptual response to intersensory discrepancy," Psychol. Bull. **88**, 638–667.
White, C. N., Ratcliff, R., and Starns, J. J. (**2011**). "Diffusion models of the flanker task: Discrete versus gradual attentional selection," Cognit. Psychol. **63**, 210–238.
Wickelgren, W. A. (**1977**). "Speed-accuracy tradeoff and information processing dynamics," Acta Psychol. (Amst.) **41**, 67–85.
Wiecki, T. V., Sofer, I., and Frank, M. J. (**2016**). "Stimulus coding with HDDMRegression — HDDM 0.6.0 documentation," http://ski.clps.brown.edu/hddm_docs/tutorial_regression_stimcoding.html (Last viewed November 8, 2016).
Wiecki, T. V., Sofer, I., and Frank, M. J. (**2013**). "HDDM: Hierarchical Bayesian estimation of the Drift-Diffusion Model in Python," Front. Neuroinformatics **7**, 14.