

# Does crossmodal correspondence modulate the facilitatory effect of auditory cues on visual search?

Anna Klapetek · Mary Kim Ngo · Charles Spence

Published online: 31 May 2012  
© Psychonomic Society, Inc. 2012

**Abstract** The “pip-and-pop effect” refers to the facilitation of search for a visual target (a horizontal or vertical bar whose color changes frequently) among multiple visual distractors (tilted bars also changing color unpredictably) by the presentation of a spatially uninformative auditory cue synchronized with the color change of the visual target. In the present study, the visual stimuli in the search display changed brightness instead of color, and the crossmodal congruency between the pitch of the auditory cue and the brightness of the visual target was manipulated. When cue presence and cue congruency were randomly varied between trials (**Experiment 1**), both congruent cues (low-frequency tones synchronized with dark target states or high-frequency tones synchronized with bright target states) and incongruent cues (the reversed mapping) facilitated visual search performance equally, relative to a no-cue baseline condition. However, when cue congruency was blocked and the participants were informed about the pitch–brightness mapping in the cue-present blocks (**Experiment 2**), performance was significantly enhanced when the cue and target were crossmodally congruent as compared to when they were incongruent. These results therefore suggest that the crossmodal congruency between auditory pitch and visual brightness can influence performance in the pip-and-pop task by means of top-down facilitation.

**Keywords** Multisensory processing · Visual search · Color · Lightness · Brightness

Although our senses are subjected to a near-constant barrage of incoming information, our brains appear to effortlessly combine these inputs into meaningful multisensory percepts representing the objects and events that fill the environments in which we live. The question of how the brain “knows” which stimuli to integrate and which to keep separate constitutes the core of the *crossmodal binding problem*, which poses a major challenge to researchers working in the area of multisensory perception (see Spence, Ngo, Lee, & Tan, 2010). Some basic principles have already been identified, such as the temporal and spatial rules (Meredith & Stein, 1986; Stein & Meredith, 1993). According to these rules, multisensory integration is more likely to occur when the constituent unimodal stimuli are co-localized in space and time (but see Spence, 2012). Additional criteria are, however, needed in order to explain why certain stimuli are integrated more efficiently than others. Recently, Spence (2011) reviewed the evidence pointing to the need to consider crossmodal correspondences as a possible additional constraint on crossmodal binding. The term “crossmodal correspondence” refers to our cognitive system’s tendency to preferentially associate certain features or dimensions of stimuli across sensory modalities. This has been demonstrated in many different studies using a variety of experimental paradigms (see Spence, 2011, for a review). While some authors have also spoken of crossmodal correspondences when they referred to explicit semantic relations between the sensory representations of the same object in different modalities (e.g., the sound of an engine and the picture of a car), here we will use the term exclusively for those correspondences between simple sensory features. The

---

A. Klapetek (✉)  
Allgemeine und Experimentelle Psychologie,  
Department Psychologie, Ludwig-Maximilians-Universität,  
Leopoldstr. 13,  
80802 München, Germany  
e-mail: anna.klapetek@psy.lmu.de

A. Klapetek · M. K. Ngo · C. Spence  
Crossmodal Research Laboratory, Department of Experimental  
Psychology, University of Oxford,  
Oxford, England, UK

experiments reported here focus on the nature (and consequences) of the crossmodal correspondences that exist between auditory pitch and the various features of visual stimuli.

One of the first systematic investigations of crossmodal correspondences was conducted by Pratt (1930). He presented tones of different pitches from a hidden loudspeaker and asked participants to indicate the vertical location from which the tones appeared to have originated, using a numerical scale arranged from floor to ceiling. The results revealed that participants assigned higher-pitched tones to higher numbers. Pratt took these results as evidence that pitch is mapped to vertical position in phenomenological space (see also Roffler & Butler, 1968; Trimble, 1934).

More recent studies have typically used variants of the speeded classification paradigm (see Garner, 1974) in which participants have to classify stimuli along one dimension, while the stimuli presented in a second, task-irrelevant stimulus dimension may or may not vary. The question addressed by researchers is whether such variation affects the latencies and, to a lesser extent, the accuracy of participants' responses. The basic idea behind such studies is that if two stimulus dimensions can be attended to independently (i.e., if they are "separable"), irrelevant variations along one dimension should not affect performance in a discrimination task based on the other dimension. By contrast, if the dimensions cannot be separated by selective attention (i.e., if they are "integral"), then effects of variations along the irrelevant dimension on response latencies should be expected.

When two stimulus dimensions are directional (e.g., when they can be classified along a common dimension, such as *high–low*, *large–small*, or *positive–negative*), their mutual relationship can be defined in terms of their congruency, such that features that share the same directional value are said to be *congruent*, whereas features with opposite directional values are considered to be *incongruent*. Many studies have demonstrated the existence of a relationship between auditory pitch and different visual features, such as spatial elevation (Ben-Artzi & Marks, 1995; Bernstein & Edelstein, 1971; Evans & Treisman, 2010; Melara & O'Brien, 1987; Patching & Quinlan, 2002; Rusconi, Kwan, Giordano, Umiltà, & Butterworth, 2006), brightness (Ludwig, Adachi, & Matzuzawa, 2011; Marks, 1987) or lightness (Hubbard, 1996; Marks, 1987; Martino & Marks, 1999; Melara, 1989; Mondloch & Maurer, 2004), size (Evans & Treisman, 2010; Gallace & Spence, 2006; Mondloch & Maurer, 2004), angularity of shape (Marks, 1987), direction of movement (Clark & Brownell, 1976), and even spatial frequency (Evans & Treisman, 2010).

So, for example, Miller (1991) used a go/no-go task in order to investigate whether crossmodal congruency modulates the *redundant-target effect* (the observation that RTs

are faster to bimodal as compared to unimodal targets; see Miller, 1982). Participants had to make a speeded response to visual or auditory targets (visual stimuli presented either above or below fixation and tones that were either high or low in frequency) while withholding their response to non-targets (visual stimuli presented at fixation or tones of intermediate pitch). The stimuli were presented either unimodally or as crossmodally congruent or incongruent bimodal stimulus pairs. The results demonstrated that participants responded significantly more rapidly to congruent than to incongruent bimodal targets. Such a result is incompatible with race models, which assume that the response is determined by the faster of the two modalities, as well as with models that assume independent coactivation of both modalities. Instead, Miller's (1991) results suggest that the redundant targets interacted in the process of evoking a behavioral response, meaning that some sort of neural integration of the auditory and visual information had taken place.

A problem both for Miller's (1991) study and for the majority of the other published studies that have utilized the speeded classification task is that they do not necessarily demonstrate whether auditory and visual stimuli are integrated at a perceptual level, or whether instead the effects of crossmodal congruency may arise during later stages of information processing—stages that may be related more to decision making and/or to response selection. For this very reason, researchers have subsequently attempted to use paradigms that allow them to rule out, or at least to minimize, the influence of decision/response selection on participants' performance.

Parise and Spence (2009), for instance, demonstrated that people are worse at judging the temporal order of two stimuli (e.g., "which stimulus came second?"), a tone or a visually presented shape, if the stimuli happen to be crossmodally congruent in terms of their auditory pitch–visual size (Experiment 1) or auditory waveform–visual shape (Experiment 2) relationship, as compared to when they happen to be crossmodally incongruent. Specifically, congruent pairs of audiovisual stimuli had to be separated by a larger temporal interval in order to be perceived as distinct perceptual events. In a third experiment, Parise and Spence (2009) investigated the effect of crossmodal congruency on the spatial ventriloquism effect. Spatial ventriloquism occurs when the apparent location of a sound is shifted in the direction of a spatially discordant visual stimulus. The participants in this study had to judge whether an auditory stimulus (a high- or low-pitched tone) had been presented to the left or the right of a simultaneously presented Gaussian blob (which was either large or small). The presentation of crossmodally congruent pairings of auditory and visual stimuli (e.g., a high tone together with a small blob or a low tone with a large blob) resulted in a significantly larger

spatial ventriloquism effect. This means that participants' discrimination thresholds were higher relative to the incongruent condition. Given that such findings cannot be explained in terms of decision- or response-related mechanisms, Parise and Spence's (2009) results seem to reflect the tendency of our cognitive systems to integrate crossmodally congruent stimuli at a perceptual level (see also Bien, ten Oever, Goebel, & Sack, 2012; Parise & Spence, 2008).

More recently, Evans and Treisman (2010) documented crossmodal congruency effects between auditory pitch and the visual dimensions of size and spatial frequency using a speeded classification task. To be able to rule out a response-related origin for these effects, they manipulated, in half of the experiments, the crossmodal correspondence between the task-irrelevant stimulus and a task-irrelevant feature of the to-be-classified stimulus (e.g., the pitch of the tone when participants had to discriminate whether it came from a violin or a piano), and additionally they ensured an orthogonal response mapping between the directly facilitated dimension (i.e., pitch, in this example) and the indirectly facilitated dimension (instrument, in this example) during the entire experiment. After having taken these steps, Evans and Treisman were able to argue that the shorter response times (RTs) in response to congruent as compared to incongruent stimulus pairs most likely reflect facilitation at perceptual stages of processing.

Note that most of the previously conducted studies on the crossmodal correspondence between auditory and visual features have focused on the effects of a task-irrelevant stimulus presented in one modality on responses to a stimulus in another sensory modality. One feature common to many of these tasks is that they were kept as simple as possible, involving the presentation of one or, at most, a few stimuli at any one time. In everyday life, however, the demands on our perceptual system are usually much higher, since auditory and visual stimuli do not appear in isolation, but typically as part of a much more complex and dynamically changing multisensory perceptual environment. If crossmodal correspondences really do play an important role in perceptual binding, as has been hinted at by some researchers (e.g., Calvert & Thesen, 2004; Spence, 2011), their effects might be expected to be most prominent in a more naturalistic context, where binding stimuli crossmodally constitutes a more challenging problem for the nervous system.

The present study was designed to investigate the frequently reported correspondence between auditory pitch and visual brightness or lightness (Hubbard, 1996; Marks, 1987; Martino & Marks, 1999; Melara, 1989; Mondloch & Maurer, 2004) under conditions in which the integration of auditory and visual stimuli would be task-relevant. We decided to use a slightly modified version of the pip-and-pop task, a visual search task originally developed by Van

der Burg, Olivers, Bronkhorst, and Theeuwes (2008) in which participants' performance is facilitated by the presence of an auditory cue synchronized with color changes of the visual target (see also Ngo & Spence, 2010; Spence & Ngo, 2012).

In Van der Burg et al.'s (2008) original study, the participants had to search for a horizontal or vertical line segment presented among distractor line segments presented in other orientations. The colors of both the target and the distractors changed regularly from red to green, or vice versa. In some blocks of trials, a short beep was presented over headphones. The onset of the auditory stimulus was synchronized with the changing of the color of the visual target. The main findings consisted of a reduction in average search latencies of more than 1,000 ms and a significant flattening of the search slope (from 147 to 31 ms/item) in cue-present relative to cue-absent trials. Importantly, a comparable facilitation of participants' visual search performance was not observed with visual cues carrying exactly the same temporal information (Van der Burg et al., 2008, Exp. 2). Moreover, the auditory cues were found to be more effective when they lagged slightly behind the visual target than when they preceded it (Van der Burg et al., 2008, Exp. 3), which is inconsistent with the typical temporal window for warning effects, and hence seems to reflect audiovisual integration instead (see Diederich & Colonius, 2008). Decreasing the temporal predictiveness of the auditory cues, by making them temporally synchronous with the target on only 20 % of the trials, did not change the overall pattern of the crossmodal facilitation effect observed in those valid trials (Van der Burg et al., 2008, Exp. 4). This result was taken to suggest that the effect was mainly stimulus-driven and automatic (cf. Lippert, Logothetis, & Kayser, 2007; Spence & Ngo, 2012). On the basis of the combined results of their four experiments, Van der Burg et al. (2008) concluded that the temporally synchronous auditory cue was perceptually integrated with the visual target, thus increasing its saliency, and hence making it "pop out" from the distractors (but see Spence & Ngo, 2012, for alternative explanations). Additional evidence for the automaticity of the pip-and-pop effect comes from a recent electroencephalographic study (Van der Burg, Talsma, Olivers, Hickey, & Theeuwes, 2011), in which an early multisensory integration effect was observed (i.e., an audiovisual event-related potential over the left parieto-occipital cortex) occurring within 50–60 ms of stimulus onset that mediated the behavioral search benefits associated with the presence of the auditory cue. What is more, this effect also occurred when the auditory cue was synchronized with a distractor, a finding that provides further support for the automaticity of the audiovisual integration process.

In the present study, we used a modified version of the original pip-and-pop paradigm, in which we manipulated

the crossmodal congruency between the pitch of the auditory cue and the brightness of the visual target. We reasoned that if the congruency between pitch and brightness were to modulate the crossmodal facilitatory effect of the cue, this could be taken as evidence for its functional role in visual search (i.e., in solving the crossmodal binding problem).

## Experiment 1

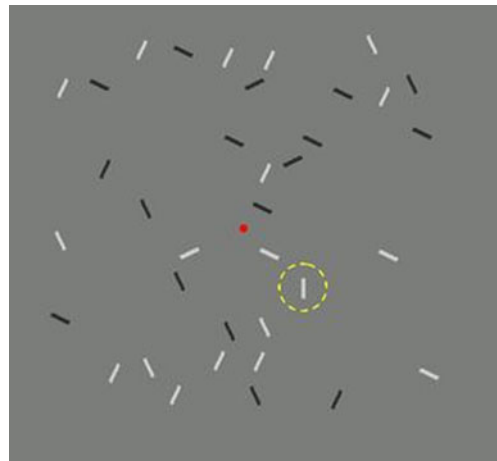
**Experiment 1** was designed to investigate whether participants' performance in the pip-and-pop task (see Ngo & Spence, 2010; Van der Burg et al., 2008) would be influenced by the crossmodal congruency between the pitch of the spatially uninformative auditory cue and the brightness of the associated visual target.

## Methods

**Participants** Sixteen participants (ten female, six male; four were left-handed), 22–46 years of age ( $M = 27.6$  years), took part in this experiment. All of the participants reported normal auditory acuity and normal or corrected-to-normal vision. The experiment lasted for approximately 45 min, and the participants received a £5 (UK Sterling) gift voucher, or course credit, in return for taking part in the study.

**Apparatus and stimuli** The participants were seated in a dimly illuminated sound-attenuated chamber in front of a 17-in. CRT monitor (screen refresh rate = 75 Hz) positioned at a viewing distance of 60 cm. The stimuli were presented using E-Prime (Psychology Software Tools, Pittsburgh, PA). The auditory stimuli consisted of either low- or high-frequency sine-wave tones (250 vs. 2000 Hz; 44.1-kHz sampling rate, 16-bit), lasting for 60 ms and presented via closed-ear headphones at approximately 60 dB. The subjective loudness of the high- and low-pitched tones was equated by reducing the amplitude of the high-pitched tone by 10 %, according to ISO 226 equal-loudness contours (see [http://en.wikipedia.org/wiki/Equal-loudness\\_contour](http://en.wikipedia.org/wiki/Equal-loudness_contour), downloaded November 2, 2010).

The visual search displays consisted of 24, 36, or 48 light gray (luminance:  $30.5 \text{ cd/m}^2$ ) or dark gray ( $0.57 \text{ cd/m}^2$ ) line segments (subtending approximately  $0.57^\circ$  of visual angle in length and  $0.08^\circ$  in width) presented against a medium gray ( $8.8 \text{ cd/m}^2$ ) background (see Fig. 1 for an example of a search display). The brightness (light or dark) of each stimulus was determined randomly at the beginning of each trial. Thus, the numbers of light and dark gray segments were uneven in most trials. The orientation of each line segment deviated randomly by exactly  $22.5^\circ$  from the horizontal or the vertical, except for the target, which was oriented either horizontally or vertically. All of the stimuli were randomly



**Fig. 1** Sample visual search display with 36 stimuli used in Experiment 1. The visual target (a vertically oriented bar) is highlighted by a dotted yellow circle (not present in the actual experiment)

placed in an invisible  $10 \times 10$  grid ( $9.58^\circ \times 9.58^\circ$ ,  $0^\circ$ – $0.34^\circ$  jitter) centered on a red fixation point, with the only restriction being that the target could never appear in one of the four central positions.

The random brightness changes of the target and distractors were achieved by generating a random sequence of different display cycles and presenting them one after the other, without interruption. One display cycle consisted of a sequence of nine display screens lasting for 50, 100, or 150 ms, each display duration being repeated randomly three times. During each display screen, some stimuli (either a certain number of the distractors or just the target) changed brightness: If the set size was 24, one, two, or three distractors changed brightness; if the set size was 36, one, three, or five distractors changed; and if the set size was 48, one, four, or seven distractors changed. Note that when the brightness of the target changed, it was the only item in the display to do so. In this case, the preceding display duration was always 150 ms, and the following display duration was always 100 ms. The brightness of the target always changed exactly once per display cycle (i.e., once every 900 ms, on average), and never during the first 500 ms of a trial. Ten different display cycles were generated per trial and presented sequentially until the participant responded or until the last display cycle had terminated, at which time the next trial started. Our design did not include target-absent trials, meaning that the target was present and also changed brightness in every trial.

**Design and procedure** During the cue-present trials, the high- and low-pitched beeps were presented in synchrony with the changing brightness of the target, so that when the target changed brightness, participants heard a beep played at one frequency and when, sometime later, the target changed back to its original brightness value, they heard



the beep at another frequency. The mapping of cue to target brightness was either congruent (a low-pitched beep was synchronized with the darker stimulus, and the high-pitched beep with the brighter stimulus) or incongruent (with the crossmodal mapping reversed). A third of the trials were cue-absent trials, in which no sounds were presented. The set size (24, 36, or 48), target orientation (horizontal or vertical), and cue condition (congruent, incongruent, or no cue) varied randomly on a trial-by-trial basis. Each of the 18 possible combinations was repeated twice per block. In total, each participant completed nine blocks of 36 trials. These test blocks were preceded by two identical practice blocks (which were not analyzed).

Each trial started with the presentation of a fixation point. After 1,000 ms the search display appeared, and it remained until the participant made a response or until all ten display cycles had been terminated. The participants were instructed to detect the target and to indicate its orientation as rapidly and accurately as possible by pressing the “z” or “m” key on the computer keyboard. The assignment of the target orientations to response keys was counterbalanced across participants. The RT reflected the time between the onset of the search display and the participant’s buttonpress. Note that while the target was present from the moment that the display was first presented, it only began to change brightness 500–900 ms later. Participants could thus detect the target before it began to change brightness, which happened in a small proportion of the trials. At the end of each block of trials, the participants received feedback concerning the mean accuracy of their responses and their RTs in the preceding block. They were allowed to take a short break, should they so desire, before starting the next block.

## Results and discussion

The mean accuracy of participants’ responding was high (all participants performed above 85 % correct; this was also the case in the subsequent visual search experiment reported here). Only correct response trials were included in the RT analysis (97 % of all trials). The data from the practice blocks and outliers (where the RT fell beyond 2.5 *SDs* from the participant’s mean in a given condition, which was the case in 3.7 % of all correct-response trials) were excluded from the analysis, as well. The mean accuracy rates and RTs for all experimental conditions in [Experiment 1](#) are summarized in [Table 1](#).

The fact that RTs were still relatively long in the tone-present condition (>2 s) may raise doubts about whether the tone could induce pop-out. Here, it is important to note that the first beep occurred 500–900 ms after display onset, and that participants may have waited for this first beep to start their search (this tendency had already been observed by Van der Burg et al., 2008, and some of our participants admitted to it as well), which means that

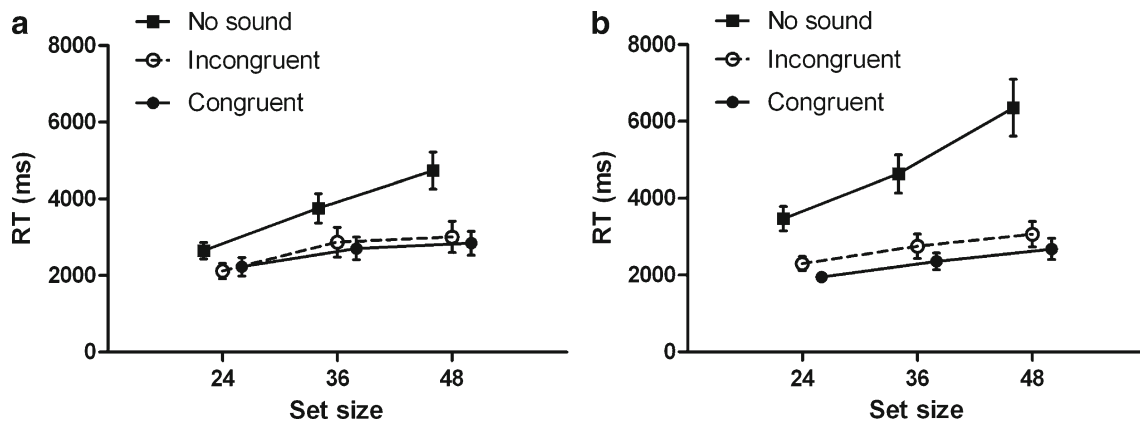
**Table 1** Mean response times (RTs, in milliseconds; numbers in parentheses represent standard errors) and accuracy (percent correct) for all combinations of set size (24, 36, and 48) and sound (congruent, incongruent, and no sound) in [Experiments 1 and 2](#)

	Set Size	Sound			Mean
		Congruent	Incongruent	No Sound	
Exp. 1	24	2229 (244) 97 %	2115 (203) 97 %	2652 (257) 97 %	2332 (201) 97 %
	36	2705 (298) 96 %	2868 (388) 97 %	3754 (401) 96 %	3109 (321) 97 %
	48	2843 (309) 97 %	3004 (411) 97 %	4739 (479) 95 %	3529 (358) 96 %
	Mean	2592 (270) 97 %	2662 (321) 97 %	3715 (339) 96 %	
Exp. 2	24	1952 (131) 95 %	2298 (183) 98 %	3473 (316) 98 %	2574 (158) 97 %
	36	2355 (222) 96 %	2752 (315) 98 %	4634 (496) 96 %	3247 (286) 97 %
	48	2679 (278) 96 %	3064 (333) 95 %	6356 (741) 94 %	4033 (344) 95 %
	Mean	2329 (198) 96 %	2705 (267) 97 %	4821 (483) 96 %	

the effective RTs can be regarded as 750 ms shorter than those reported here.

A repeated measures analysis of variance (ANOVA) was conducted on the RT data with Sound (congruent, incongruent, or no sound) and Set Size (24, 36, or 48) as the within-participants factors. The Huynh–Feldt correction was applied whenever sphericity could not be assumed. This analysis revealed significant main effects of sound,  $F(1.2, 18.1) = 17.37, p < .001, \eta^2 = .54$ , and set size,  $F(1.5, 22.5) = 27.89, p < .001, \eta^2 = .65$ . We also found a significant interaction between these two factors,  $F(4, 60) = 8.82, p < .001, \eta^2 = .37$ . The RTs were similar in the congruent ( $M = 2,592$  ms) and the incongruent ( $M = 2,662$  ms) conditions and were substantially longer when no sound was presented ( $M = 3,715$  ms). Pairwise comparisons (Tukey’s HSD) confirmed that the mean RTs differed significantly between the no-sound and congruent-sound conditions ( $p < .001$ ) and between the no-sound and incongruent-sound conditions ( $p = .001$ ). There was no significant difference between the congruent- and incongruent-sound conditions. As expected, RTs increased as the set size increased (at set size 24,  $M = 2,332$  ms; at set size 36,  $M = 3,109$  ms; and at set size 48,  $M = 3,529$  ms). The main effects of sound and set size, as well as their interaction, can be clearly seen in [Fig. 2a](#).

A second repeated measures ANOVA conducted on the individual participants’ search slopes also revealed a significant main effect of sound,  $F(2, 30) = 16.01, p < .001, \eta^2 = .52$ . Post-hoc comparisons revealed that the slope in the cue-absent



**Fig. 2** Response times (RTs) as a function of sound (congruent, incongruent, or no sound) and set size (24, 36, or 48). Error bars represent the standard errors of the means for each combination of the two factors in [Experiments 1](#) (panel a) and [2](#) (panel b)

condition (87 ms/item) was significantly steeper than those in both cue-present conditions ( $p < .001$  for the comparison to the congruent condition, and  $p = .002$  for the comparison to the incongruent condition). Although the slope in the incongruent condition (37 ms/item) was steeper than that in the congruent condition (26 ms/item), this difference was not statistically significant.

A similar analysis of the error data revealed no significant main effects, nor any interaction between the two factors Sound and Set Size. What is more, the comparison of the participants' performance in the correct- and incorrect-response trials revealed no evidence for a speed-accuracy trade-off.

In summary, three main findings emerged from the analysis of the results of [Experiment 1](#): First, we successfully replicated the pip-and-pop effect (Ngo & Spence, 2010; Van der Burg et al., 2008). That is, participants' visual search performance was significantly facilitated (in terms of faster RTs and reduced slopes) by the presence of an auditory cue synchronized with a change in the brightness of the visual target, hence demonstrating the robustness of this particular crossmodal effect.

Second, the results also demonstrated that the findings of previous studies that have used the pip-and-pop task (Ngo & Spence, 2010; Van der Burg et al., 2008) can be extended to the situation in which the targets change in brightness rather than hue (note that, in all previous studies of the pip-and-pop task, the targets changed from red to green). Third, and contrary to our predictions, the crossmodal correspondence between the pitch of the accessory sound and the brightness of the visual target did not modulate participants' performance in this task when the congruency of the auditory accessory stimulus and the visual target varied on a trial-by-trial basis.

In a recent study by Van der Burg, Cass, Olivers, Theeuwes, and Alais (2010), the presence of an amplitude-modulated auditory signal was found to facilitate

participants' search for a visual target that periodically changed luminance with the same frequency, independent of whether the signals were presented in phase or counter-phase. However, before we compare our study to theirs, we should emphasize that the two studies differ in two key respects: First, Van der Burg et al. (2010) manipulated the correspondence between loudness and brightness, while we controlled the correspondence between pitch and brightness instead. Second, the temporal modulation in their study was strictly periodic, which was not the case in the original pip-and-pop task, in which the durations of the color states were randomly determined as each trial proceeded. However, the reason why we discuss Van der Burg et al.'s (2010) recent findings here relates to their main conclusion: namely, that crossmodal facilitation occurs when the cognitive system detects synchronous audiovisual transients, with the direction of the transients playing no role at all. When an auditory onset is synchronized with a visual offset, the brain still perceives a simultaneous audiovisual transient event. If transients really do play the essential role in the pip-and-pop experiment, as Van der Burg et al. (2010) suggested, the effect of such transients on performance might be expected to override any potential influence of the crossmodal congruency between the temporally modulated auditory and visual features. This was exactly the case in [Experiment 1](#), in which congruent and incongruent cues were equally effective in facilitating participants' visual search performance. The results of [Experiment 1](#) therefore suggest that the crossmodal congruency between the pitch of a spatially nonpredictive auditory cue and the brightness of a visual target does not affect performance in the pip-and-pop task, at least not in an automatic manner. However, it is important to note that this does not necessarily rule out any potential effects of this type of crossmodal correspondence on performance in the visual search task; it may simply mean that information about crossmodal congruency needs to be

processed consciously/deliberately. In [Experiment 1](#), cross-modal congruency was varied on a trial-by-trial basis. As a result, it seems possible that the participants may simply not have noticed the difference between congruent and incongruent trials, perhaps because they simply did not pay any attention to the pitch of the cues. In [Experiment 2](#), the participants were therefore made aware of the crossmodal congruency, and congruency was varied on a block-by-block (rather than trial-by-trial) basis. The hope was that these methodological changes might result in improved visual search performance by participants in the crossmodally congruent, as opposed to the incongruent, blocks of trials. Because we noticed that participants spontaneously referred to the stimuli as “black” and “white” instead of “dark” and “bright,” we decided to label the dark stimulus “black” and the bright stimulus as “white” (although their RGB values corresponded to very dark and very light gray, respectively).

## Experiment 2

### Method

Sixteen participants (11 female, five male; all right-handed) 17–38 years of age ( $M = 28.6$  years) took part in [Experiment 2](#). All of the participants reported normal auditory acuity and normal or corrected-to-normal vision. The experiment lasted for approximately 45 min, and the participants received a £5 gift voucher or course credit in return for taking part in the study.

The apparatus and materials were exactly the same as in [Experiment 1](#). The design, however, was modified so that the crossmodal congruency between the pitch of the sound and the brightness of the visual target was varied on a block-by-block basis (rather than on a trial-by-trial basis, as in [Experiment 1](#)). Three blocks of trials were congruent (with the low-pitched beep being synchronized with the low-luminance states of the target, whereas the high-pitched beep was synchronized with the high-luminance states of the target), three were incongruent (using the reverse mapping), and in three baseline blocks, no sounds were presented. At the beginning of each block of trials, the participants were informed about the pitch–color mapping between the auditory cue and the visual target and were instructed to try and make use of this information. The set size (24, 36, or 48) and target orientation (horizontal or vertical) varied randomly on a trial-by-trial basis, with each of the six combinations being repeated six times per block. In total, nine blocks of 36 trials each were presented. These test blocks were preceded by three shorter practice blocks (18 trials each), one for each cue condition. The order in which the congruent (C), incongruent (I), and no-sound (N) conditions were presented was completely counterbalanced across participants, meaning that each

participant was systematically assigned one of the six possible orders (CIN, CNI, ICN, INC, NCI, and NIC), which was then repeated four times (once during practice and three times during the experiment).

### Results and discussion

Only correct-response trials (96.3 % of all trials) were included in the analysis of the RT data. The data from the practice blocks and from outliers (in which the RTs fell beyond 2.5 *SDs* from the participant’s mean in a given condition, which was the case in 3.7 % of all correct-response trials) were excluded from the analysis as well.

A repeated measures ANOVA was conducted on the RT data with Sound (congruent, incongruent, or no sound) and Set Size (24, 36, or 48) as the within-participants factors. The Huynh–Feldt correction was applied whenever sphericity could not be assumed. Analysis of the RT data revealed significant main effects of sound,  $F(1.1, 17) = 23.96$ ,  $p < .001$ ,  $\eta^2 = .62$ , and set size,  $F(2, 30) = 33.41$ ,  $p < .001$ ,  $\eta^2 = .69$ . There was also a significant interaction between these two factors,  $F(2.1, 31.9) = 10.08$ ,  $p < .001$ ,  $\eta^2 = .40$ . Pairwise comparisons (Tukey’s HSD) confirmed that the mean RTs differed significantly between the no-sound and congruent-sound conditions (mean difference of 2,492 ms,  $p < .001$ ), between the no-sound and incongruent-sound conditions (mean difference of 2,116 ms,  $p < .001$ ), and, importantly, also between the congruent- and incongruent-sound conditions (mean difference of 376 ms,  $p = .009$ ).

A second repeated measures ANOVA conducted on the search slopes of individual participants also revealed a significant main effect of sound,  $F(1.3, 19.7) = 12.60$ ,  $p = .001$ ,  $\eta^2 = .46$ . Post-hoc comparisons (Tukey’s HSD) confirmed that the mean search slope in the cue-absent condition (120 ms/item) was significantly steeper than those in both the congruent ( $p = .002$ ) and the incongruent ( $p = .003$ ) conditions. However, the search slopes in the congruent (30 ms/item) and incongruent (32 ms/item) conditions were nearly identical (see [Fig. 2b](#)). The difference between the latter two conditions failed to reach statistical significance, meaning that one cannot conclude that visual search was any more efficient in those blocks of trials in which the auditory cue happened to be crossmodally congruent with the target than in those blocks in which it happened to be incongruent.

A third ANOVA performed on the error data documented no significant main effects of sound or set size, nor any interaction between these two factors (just as had been the case in [Experiment 1](#)). Once again, comparison of the RTs from the correct- and incorrect-response trials revealed no evidence of a speed–accuracy trade-off in participants’ performance. The mean RTs and accuracy rates for all experimental conditions in [Experiment 2](#) are summarized in [Table 1](#).

The results of [Experiment 2](#) therefore demonstrate that when participants knew the crossmodal mapping between the auditory and visual features in advance, they responded to visual targets paired with crossmodally congruent sounds significantly more rapidly than when the targets were paired with an incongruent sound. This result, together with the fact that crossmodal congruency had had no effect when congruent and incongruent trials were presented unpredictably on a trial-by-trial basis (in [Experiment 1](#)), suggests that the particular crossmodal correspondence between auditory pitch and visual brightness<sup>1</sup> operates at a more strategic (i.e., rather than at an automatic or involuntary) level.

The experimental design used in the present study did not involve a condition with a neutral cue. Consequently, we cannot say for sure whether the RT difference reported here should be attributed to facilitation in the crossmodally congruent blocks of trials, interference in the crossmodally incongruent blocks of trials, or some unknown combination of the two effects (cf. [Jonides & Mack, 1984](#)). Nor can we assert unequivocally whether the modulation of crossmodal congruency affects participants' perceptual, decisional, and/or response-related processes. The fact that we observed an effect of crossmodal congruency in [Experiment 2](#) and not in [Experiment 1](#) suggests that our participants used some sort of top-down strategy. It is known that search for a target defined by the conjunction of color and a second feature is more efficient when observers know the color in advance, as they can limit their search to the subset of display elements defined by that color ([Wolfe, Cave, & Franzel, 1989](#); see also [Kaptein, Theeuwes, & Van der Heijden, 1995](#)). This may have been the case in [Experiment 2](#), in which participants may have used their knowledge about the given pitch–color mapping in order to limit their search to the color indicated by the tone at a given moment (or for a given block of trials). It is reasonable to assume that semantically congruent cues (i.e., low tones paired with dark targets and high tones paired with bright targets) gave rise to a more efficient (in terms of higher accuracy and resistance to interference at a lower computational cost) working memory representation of the target color, and consequently may have allowed for more efficient attentional guidance. Although the results strongly suggest that congruency operated through the top-down strategic use of the cues, we should still consider the possibility that congruency modulated the strength of the crossmodal binding between the cue and the target, and thereby directly influenced the target's perceptual saliency. It is very likely that the instructions in [Experiment 2](#) forced participants to allocate more attention to the processing of the auditory stimuli, which

may have facilitated their perceptual integration with the target. This would be consistent with other evidence that audiovisual integration requires attentional resources (see, e.g., [Alsius, Navarra, Campbell, & Soto-Faraco, 2005](#); [Fairhall & Macaluso, 2009](#); [Talsma, Doty, & Woldorff, 2007](#)).

The question of whether congruency operated automatically or through the strategic use of the cues by participants can best be answered on the basis of an analysis of the time course of the congruency effect. If congruency plays a role in crossmodal binding, its effect should be observable at the moment when the target pops out, which in most trials happens on the first or the second beep ([Van der Burg et al., 2008](#)). In contrast, if congruency has no influence on the perceptual saliency of the target, and instead modulates visual search only by means of some kind of top-down influence, its effects should manifest at a later point in time, given that participants would most likely adopt the top-down strategy on those trials in which the target did not (for whatever reason) pop out. To see at which point in time congruency showed the largest effect, we therefore compared the frequency distributions of RTs in congruent, incongruent, and cue-absent trials.<sup>2</sup> In line with [Van der Burg et al. \(2008\)](#), we adjusted the raw RTs by subtracting the random period of 500–900 ms that preceded the first target color change (and the first onset of the tone, in the tone-present trials), and thus only considered the RT from the moment that the cue had been present. To do so, we had to discard the 4 % of all trials, in which the participants responded to the target before the onset of the cue.

#### Analysis of RT distributions

[Figure 3](#) shows the distributions of the adjusted RTs for congruent, incongruent, and cue-absent trials, pooled across all set sizes and time-locked to the first target change (which corresponded with the time of the first beep in the tone-present conditions).

In [Experiment 1](#), the distribution of RTs from the tone-absent trials displays a peak around 1,200 ms, which corresponds, on average, to the time after the second target-color switch, indicating that in a substantial proportion of the trials, the participants detected the target after it had changed color only once or twice. The tone-present distributions show a much higher peak at around 1,400 ms, which corresponds, on average, to the time between the second and third target changes and auditory signals. As can be seen in the upper right panel of [Fig. 3](#), where the relative proportions of short, medium, and long RTs are plotted as a function of cue condition, there is a clear shift toward short RTs in tone-present as compared to tone-absent trials, which is consistent with the findings of [Van der Burg et al. \(2008\)](#). This result suggests that in a substantial proportion of the trials,

<sup>1</sup> Although, in everyday language, black and white are classified as colors, most people are aware that these “colors” have no particular hue, but rather correspond to the two poles of the brightness dimension.

<sup>2</sup> We thank Eric Van der Burg for making this suggestion.



the presence of the tone led to fast and automatic detection of the target. For the purposes of the main research question addressed here—namely, whether and how pitch–brightness congruency modulates the pip-and-pop effect—the crucial finding is that the congruent and incongruent distributions are nearly identical, which confirms our previous conclusion that crossmodal congruency had no effect on visual search performance in [Experiment 1](#).

In [Experiment 2](#), all three distributions peak at about the same time (1,100–1,200 ms after the first target change and, if present, tone onset), but the difference between the tone-present and tone-absent maxima is more pronounced than in [Experiment 1](#), suggesting that the auditory cues induced “pop-out” more effectively. Before we speculate on the possible mechanisms underlying this difference, it should be noted that each experiment was conducted with a different group of participants and that the two groups may not have been entirely balanced with respect to their ability to benefit from the auditory cues (while some participants reported that the tones made search easier, others did not perceive any benefit). Overall, the mean RTs in [Experiment 2](#) were longer than those in [Experiment 1](#), but the difference was most pronounced for cue-absent trials, which indicates that in [Experiment 2](#), participants relied more on the auditory cues.

At first glance, the distributions for congruent and incongruent trials look very similar in the right lower panel of [Fig. 3](#), but a closer inspection of the proportions of short, medium, and long RTs in both conditions reveals a shift toward short RTs for the congruent condition (or a shift toward longer RTs in the incongruent condition). A shift in the same direction, although much smaller, is also observable in the data from [Experiment 1](#) (upper right panel). This result confirms that—particularly in [Experiment 2](#), in which participants knew the pitch–brightness mapping in a given block—congruency led to a higher proportion of trials in which the target attracted attention due to its saliency (we prefer not to use the term “pop out” here, since it may imply that we are referring to a purely bottom-up mechanism, which is certainly not the case).

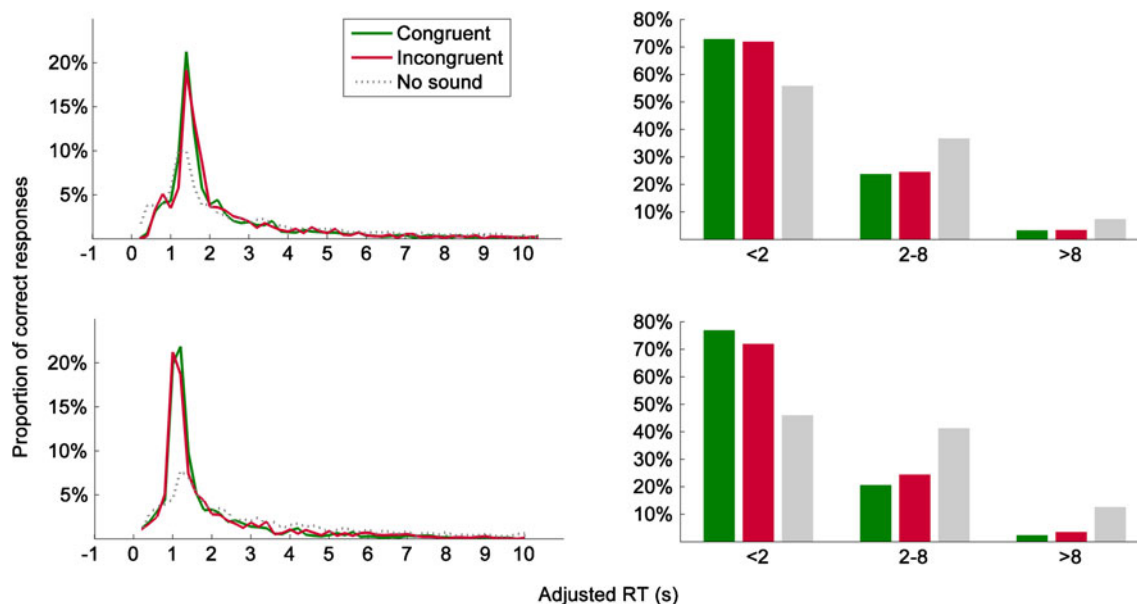
## General discussion

The results of the two experiments reported here support a number of general conclusions. First, this study represents an independent replication of the pip-and-pop effect, originally reported by [Van der Burg et al. \(2008\)](#). Following [Van der Burg et al. \(2008\)](#)’s study and that of [Ngo and Spence \(2010\)](#), in which stimuli that changed in hue were used, we used stimuli changing in luminance and also found a significant RT benefit in cue-present as compared to cue-absent trials. This finding can therefore be taken as evidence to further support the robustness and generalizability of the pip-and-pop effect.

Contrary to our expectations, though, the results of [Experiment 1](#) revealed that the crossmodal correspondence between auditory pitch and visual brightness—the existence of which has been demonstrated by a number of researchers in adults, young children, and even chimpanzees (e.g., [Hubbard, 1996](#); [Ludwig et al., 2011](#); [Marks, 1987](#); [Martino & Marks, 1999](#); [Melara, 1989](#); [Mondloch & Maurer, 2004](#))—did not influence participants’ performance on the pip-and-pop task, at least not when the congruency of the crossmodal correspondence (i.e., congruent vs. incongruent) was varied randomly on a trial-by-trial basis. The results of [Experiment 1](#) therefore seem to corroborate the view that it is mainly the temporal information concerning the auditory signal, and not some other quality, such as its pitch, that is integrated with the visual signal. Nonetheless, this does not necessarily mean that the pairing between the pitch of the cue and target luminance is irrelevant: When participants were informed of the pitch–luminance pairing utilized in a given block of trials, they performed significantly better when dark objects were synchronized with low-pitched tones and light objects with high-pitched tones ([Experiment 2](#)). These results therefore suggest that the participants could use the information about the pitch–luminance correspondence in a top-down manner and that they were better in doing so when the pairing was congruent than when it was incongruent.

A comparison of the frequency distributions of RTs from the congruent and incongruent trials in [Experiments 1 and 2](#) revealed that the proportion of very short RTs was higher in the congruent condition. This result suggests that at least part of the total congruency effect operated by means of a fast, automatic process. On the other hand, crossmodal congruency does not seem to have affected the efficiency of visual search (i.e., the search slopes were nearly identical in the crossmodally congruent and incongruent conditions). This, in turn, suggests that most of the crossmodal congruency effect arose at a postperceptual stage of information processing. It is likely that participants could respond to the target faster when they simultaneously heard a crossmodally congruent tone. They may also have experienced more ambiguity when responding while the tone was incongruent. Similar audiovisual response-compatibility effects have now been reported by a number of other researchers (e.g., [Evans & Treisman, 2010](#); [Melara & Marks, 1990](#); [Rusconi et al., 2006](#)). However, both these other researchers’ results and our own data from a control experiment<sup>3</sup> have shown that

<sup>3</sup> In order to obtain a direct measure of the semantic association between pitch and brightness, we conducted a simplified version of the Implicit Association Test ([Greenwald, McGhee, & Schwartz, 1998](#)), a method that is thought to assess stimulus–response compatibility ([De Houwer, 2003](#)), with low and high tones and black and white circles as the stimuli. Nine out of 12 participants showed a “moderate” ( $D > 0.35$ ) or “strong” ( $D > 0.65$ ) IAT effect ( $D$  is an individual effect size calculated in the IAT scoring procedure; see [Greenwald, Nosek, & Banaji, 2003](#), for a description of the algorithm).



**Fig. 3** Response time (RT) distributions in [Experiments 1](#) (upper panels) and [2](#) (lower panels). The proportions of correct responses (RT < 10 s) are plotted as a function of the adjusted RT (the bin width is 200 ms), separately for congruent, incongruent, and tone-absent trials (left panels). The adjusted RT is the time to respond to the target

from the first target-color change (and the onset of the tone, in tone-present trials). The right panels show the proportions of congruent, incongruent, and tone-absent trials with short (<2 s), medium (2 to 8 s), and long (>8 s) RTs

such effects tend to be much smaller (<100 ms) than the congruency effect observed in [Experiment 2](#) (376 ms). This leads us to the conclusion that the RT difference in [Experiment 2](#) must reflect the consequences of congruency operating at several different stages of cognitive processing.

First, it is possible that a small part of the effect is automatic: The tone may have boosted target saliency only when the pitch-brightness relationship between the two stimuli was congruent. Although this account is speculative, it is nevertheless consistent with the finding that the data from [Experiment 1](#) revealed a nonsignificant trend toward shallower search slopes in the congruent condition, which suggests that search efficiency was slightly greater with crossmodally congruent tones.

Second, crossmodal congruency may have affected participants' visual search by means of a top-down influence when the cues were used strategically by the participants. The fact that top-down influences may play some role in the pip-and-pop task has already been acknowledged by Van der Burg and his colleagues. They found, for instance, that the validity of the auditory cues (i.e., whether the onset of the beep was synchronized with the target or with a distractor in a majority of trials) modulated their effectiveness (see Van der Burg et al., 2008, Exp. 4). Such findings indicate that at least a part of the pip-and-pop effect could be explained by means of a strategic use of the tones. This is also consistent with the finding that search slopes are never flat in the tone-present condition, as one certainly

might expect to be the case if the target popped out in every trial.

Last, but not least, crossmodal congruency most likely influenced the speed of responses once the target had been detected (given its large effect on RTs but its nonsignificant effect on the efficiency of participants' visual search). To the best of our knowledge, this is the first demonstration of an implicit stimulus–response compatibility effect between auditory pitch and visual brightness. With that said, comparable effects have been demonstrated for pitch and other visual dimensions (e.g., Evans & Treisman, 2010; Melara & Marks, 1990; Rusconi et al., 2006).

All in all, the results of the present study appear to support a model in which the crossmodal congruency between the pitch of the auditory cue and the luminance of the visual target in the version of the pip-and-pop task utilized here can be represented at various levels of cognitive processing, depending on the task instructions/requirements, the strategy used by the participants, and their degree of awareness of the crossmodal correspondence. Chiou and Rich (2012), for instance, recently demonstrated that the crossmodal correspondence between auditory pitch and spatial elevation can manifest at an attentional level of processing, where its effects are susceptible to contextual manipulations and volitional control. Some support for a flexible “multi-stage” model of cross-modal integration comes from a recent magnetic encephalography study by Diaconescu, Alain, and McIntosh (2011). Their results showed that while audiovisual multisensory facilitation was associated with posterior parietal

activity as early as 100 ms after stimulus onset, multisensory processes extended to cingulate, temporal, and prefrontal cortices when the participants were required to evaluate cross-modal stimuli on the basis of their semantic category or degree of semantic congruency.

The question of why we found no significant evidence in favor of a genuinely perceptual effect of the pitch–brightness correspondence on participants’ performance is not easy to answer. It is possible that crossmodal correspondences may differ in terms of their neural origin and, accordingly, manifest at different levels of the cognitive system (see also Sadaghiani, Maier, & Noppeney, 2009). Spence (2011) suggested that crossmodal correspondences can be classified into at three distinct types: structural, which most likely reflect direct correspondences in the neural processing of sensory information; statistical, which reflect crossmodal associations between sensory features or dimensions that exist in nature (e.g., the fact that small objects make higher-pitched sounds) and are, most likely, simply learned; and semantic, which apply when two dimensions overlap in the meaning (or associations) of the stimuli. If the correspondence between auditory pitch and visual brightness were to be semantic in nature, it could possibly never manifest itself at a preattentive stage of perceptual processing (see also Gallace & Spence, 2006), while a structural or statistical correspondence might (see Ernst, 2007; Parise & Spence, 2009). However, mounting evidence is showing that the crossmodal correspondence between pitch and brightness cannot be purely semantic in nature, since it has been observed in both prelinguistic children (Mondloch & Maurer, 2004) and, more recently, in chimpanzees (Ludwig et al., 2011). On the other hand, Mondloch and Maurer (2004; Maurer & Mondloch, 2005) have hypothesized that the correspondence in adults may be a remainder of cross-modal neural connections that may have been present at birth and influenced the organization of perception and language before they were pruned out or placed under inhibitory control in the course of development. Such an account theoretically allows the crossmodal association to be semantic in adults and at the same time to be present in small children and animals.

But, even if the correspondence between pitch and brightness is structural—that is, hardwired into the neural system—it may not have manifested itself in the present study, as it is probable that not only the type of crossmodal correspondence, but also the nature of the task involved, determines the stage of cognitive processing at which any particular crossmodal correspondence will be expressed behaviorally. Considering the evidence that attention is sometimes essential for crossmodal binding (see, e.g., Alsius et al., 2005; Fairhall & Macaluso, 2009; Talsma et al., 2007), it may be that those tasks in which the attentional focus is rather narrow from the very beginning, such as simple

detection tasks or the ventriloquism paradigm, favor early effects of crossmodal correspondences, while visual search paradigms, in which attention typically has a broad focus at the start of a trial, offer little possibility for such effects. It is even possible that the lack of early congruency effects is specific to the pip-and-pop paradigm. As we have mentioned already while discussing the results of Experiment 1, it is quite probable that crossmodal congruency cannot influence perception automatically in the pip-and-pop task, because the perceptual system prioritizes the detection of audiovisual transients over the perception of the direction of these transients.

Although we suspect that the lack of any early crossmodal congruency effect was task-dependent, we cannot assert such a claim with any degree of certainty, and thus we have no grounds on which to draw firm conclusions concerning the nature of the crossmodal correspondence between auditory pitch and visual brightness. Nevertheless, we will briefly review the extant evidence for and against each of the different possibilities. One frequently mentioned argument here is that humans acquire crossmodal correspondences as they observe them (or, rather, statistical correlations) in their everyday environments. Certain crossmodal correspondences can indeed be directly linked to physical laws, and thus are frequently found in nature. So, for example, the statistical correspondence between pitch and size can be attributed to the fact that larger objects resonate at lower frequencies (e.g., Kunkler-Peck & Turvey, 2000; Lakatos, McAdams, & Caussé, 1997). However, unlike in the case of pitch and size, there is no obvious physical, nor statistical, relation between pitch and brightness in the real world. The correspondence between pitch and brightness is therefore unlikely to have been learned from observing the statistical properties of the environment. Unlike other crossmodal associations, it cannot be attributed to intensity-matching, either, since pitch is a metathetic dimension according to Stevens’ psychophysical criteria—that is, it is qualitative, not quantitative (see Stevens, 1957). Despite its absence in the natural environment, this crossmodal association (or correspondence) clearly forms part of our semantic repertoire, which is reflected in our everyday use of language, such as when we refer to high- or low-pitched sounds as “bright” or “dark,” respectively. When asked about this, people are usually aware of the correspondence, but they do not see any logic behind it, nor can they explain where it comes from. A question that will perhaps cross the reader’s mind at this point is whether a crossmodal association that does not reflect any properties of the environment has any adaptive value, and thus a biological reason to exist. For now, we can only speculate that the two dimensions share meaning on some more basic sensory or semantic dimension, such as “good–bad” or “safe–dangerous.” Certain crossmodal associations have been linked previously to the limbic system, both in humans (Cytowic, 1989) and in monkeys (e.g., Murray &

Mishkin, 1985), so it could theoretically be possible that at least some types of correspondence represent remnants of evolutionarily older forms of perception, in which senses were less under cortical control and not that well differentiated.

The fact that the correspondence between pitch and brightness neither is statistical nor can be attributed to some semantic origin suggests that it may have evolved earlier than language. Further support for this notion comes from experiments, in which this crossmodal correspondence influenced the behavior of prelinguistic children (Mondloch & Maurer, 2004), and even of chimpanzees (Ludwig et al., 2011).

According to Mondloch and Maurer (2004; Maurer & Mondloch, 2005), the correspondence is a remnant of a transient state of neonatal synesthesia as a part of normal development, during which all of the senses are linked. The authors describe several possible scenarios how this could be established at a neural level, but the core assumption is that the synesthetic connections or networks are either pruned out or placed under inhibitory control during later development (Maurer & Mondloch, 2005). What is important is that these connections or the accompanying synesthetic perceptual experiences may influence the child's developing language and semantic repertoire. A similar argument, but from a phylogenetic perspective, was put forward by Ramachandran and Hubbard (2001), who suggested that crossmodal correspondences between sensory dimensions may have “boot-strapped” the evolution of language. The notion that a neural link underlies the crossmodal correspondence between auditory pitch and visual brightness, at least at some point in development, is also supported by the fact that in the rare condition of colored-hearing synesthesia, in which auditory percepts are automatically accompanied by percepts of color, high-frequency sounds typically produce brighter percepts than do low-frequency sounds (Marks, 1974). In line with the neonatal hypothesis, many researchers believe that synesthesia results from insufficient pruning (Baron-Cohen, 1996) or deficient inhibition (Grossenbacher & Lovelace, 2001) of neural connections.

As a whole, the above evidence suggests that a classification of crossmodal correspondences into categories such as “structural” or “semantic” (in the terms of Spence, 2011) may not be too useful, since the categories do not seem to be exclusive, and their “hard-wiredness” may rather be a matter of degree.

## Conclusions

The results of the present study provide evidence that the crossmodal association (or correspondence) between auditory pitch and visual brightness modulates participants' performance in complex tasks that require the perceptual integration of auditory and visual signals, as is the case in

the pip-and-pop task. Contrary to our initial predictions, though, the results reported here do not provide any clear evidence that this particular crossmodal association plays a role in the preattentive binding of auditory and visual signals, at least under crowded conditions of stimulus presentation. Nonetheless, the finding that the crossmodal correspondence between auditory and visual signals can modulate visual search performance and can induce response-compatibility effects when the auditory and visual features are attended to certainly has a number of important implications for both theory and practice.

For instance, researchers concerned with the integration of auditory and visual information should pay particular attention to possible congruency relationships between both sensory modalities, since ignorance in this respect could introduce substantial bias into their results. In applied ergonomics, the careful consideration of crossmodal correspondences in the design of audiovisual information displays or user interfaces can greatly improve the efficiency of these devices (see, e.g., Spence et al., 2010). Future research will hopefully take up some of the problems and unanswered questions discussed here and provide further insights into the exact mechanisms underlying crossmodal correspondences between auditory and visual features, as well as into their role in audiovisual binding.

## References

- Alsius, A., Navarra, J., Campbell, R., & Soto-Faraco, S. (2005). Audiovisual integration of speech falters under attention demands. *Current Biology*, *15*, 839–843.
- Baron-Cohen, S. (1996). Is there a normal phase of synaesthesia in development? *Psyche*, *2*, retrieved on March 14, 2012, from <http://daysyn.com/BaronCohen1996.pdf>
- Ben-Artzi, E., & Marks, L. E. (1995). Visual–auditory interaction in speeded classification: Role of stimulus difference. *Perception & Psychophysics*, *57*, 1151–1162. doi:10.3758/BF03208371
- Bernstein, I. H., & Edelman, B. A. (1971). Effects of some variations in auditory input upon visual choice reaction time. *Journal of Experimental Psychology*, *87*, 241–247.
- Bien, N., ten Oever, S., Goebel, R., & Sack, A. T. (2012). The sound of size: Crossmodal binding in pitch-size synesthesia: A combined TMS, EEG and psychophysics study. *NeuroImage*, *59*, 663–672. doi:10.1016/j.neuroimage.2011.06.095
- Calvert, G., & Thesen, T. (2004). Multisensory integration: Methodological approach and emerging principles in the human brain. *The Journal of Physiology*, *98*, 191–205.
- Chiou, R., & Rich, A. N. (2012). Cross-modality correspondence between pitch and spatial location modulates attentional orienting. *Perception*, *41*, 339–353.
- Clark, H. H., & Brownell, H. H. (1976). Position, direction, and their perceptual integrality. *Perception & Psychophysics*, *19*, 328–334. doi:10.3758/BF03204238
- Cytowic, R. E. (1989). *Synesthesia: A union of the senses*. New York, NY: Springer.
- De Houwer, J. (2003). A structural analysis of indirect measures of attitudes. In J. Musch & K. C. Klauer (Eds.), *The psychology of*



- evaluation: *Affective processes in cognition and emotion* (pp. 219–244). Mahwah, NJ: Erlbaum.
- Diaconescu, A. O., Alain, C., & McIntosh, A. R. (2011). The co-occurrence of multisensory facilitation and cross-modal conflict in the human brain. *Journal of Neurophysiology*, *106*, 2896–2909.
- Diederich, A., & Colonius, H. (2008). Crossmodal interaction in saccadic reaction time: Separating multisensory from warning effects in the time window of integration model. *Experimental Brain Research*, *186*, 1–22.
- Ernst, M. O. (2007). Learning to integrate arbitrary signals from vision and touch. *Journal of Vision*, *7*(5), 1–14. doi:10.1167/7.5.7
- Evans, K. K., & Treisman, A. (2010). Natural cross-modal mappings between visual and auditory features. *Journal of Vision*, *10*(1), 1–12. doi:10.1167/10.1.6
- Fairhall, S. L., & Macaluso, E. (2009). Spatial attention can modulate audiovisual integration at multiple cortical and subcortical sites. *European Journal of Neuroscience*, *29*, 1247–1257. doi:10.1111/j.1460-9568.2009.06688.x
- Gallace, A., & Spence, C. (2006). Multisensory synesthetic interactions in the speeded classification of visual size. *Perception & Psychophysics*, *68*, 1191–1203. doi:10.3758/BF03193720
- Garner, W. R. (1974). *The processing of information and structure*. Potomac, MD: Erlbaum.
- Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. K. (1998). Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology*, *74*, 1464–1480.
- Greenwald, A. G., Nosek, B. A., & Banaji, M. R. (2003). Understanding and Using the Implicit Association Test: I. An Improved Scoring Algorithm. *Journal of Personality and Social Psychology*, *85*, 197–216.
- Grossenbacher, P. G., & Lovelace, C. T. (2001). Mechanisms of synesthesia: Cognitive and physiological constraints. *Trends in Cognitive Sciences*, *5*, 36–41. doi:10.1016/S1364-6613(00)01571-0
- Hubbard, T. L. (1996). Synesthesia-like mappings of lightness, pitch, and melodic interval. *The American Journal of Psychology*, *109*, 219–238.
- Jonides, J., & Mack, R. (1984). On the cost and benefit of cost and benefit. *Psychological Bulletin*, *96*, 29–44. doi:10.1037/0033-2909.96.1.29
- Kaptein, N. A., Theeuwes, J., & Van der Heijden, A. H. C. (1995). Search for a conjunctively defined target can be selectively limited to a color-defined subset of elements. *Journal of Experimental Psychology: Human Perception and Performance*, *21*, 1053–1069.
- Kunkler-Peck, A. J., & Turvey, M. T. (2000). Hearing shape. *Journal of Experimental Psychology: Human Perception and Performance*, *26*, 279–294.
- Lakatos, S., McAdams, S., & Caussé, R. (1997). The representation of auditory source characteristics: Simple geometric form. *Perception & Psychophysics*, *59*, 1180–1190. doi:10.3758/BF03214206
- Lippert, M., Logothetis, N. K., & Kayser, C. (2007). Improvement of visual contrast detection by a simultaneous sound. *Brain Research*, *1173*, 102–109.
- Ludwig, V. U., Adachi, I., & Matzuzawa, T. (2011). Visuoauditory mappings between high luminance and high pitch are shared by chimpanzees (*Pan troglodytes*) and humans. *Proceedings of the National Academy of Sciences*, *108*, 20661–20665.
- Marks, L. E. (1974). On associations of light and sound: The mediation of brightness, pitch and loudness. *The American Journal of Psychology*, *87*, 173–188.
- Marks, L. E. (1987). On cross-modal similarity: Auditory–visual interactions in speeded discrimination. *Journal of Experimental Psychology: Human Perception and Performance*, *13*, 384–394. doi:10.1037/0096-1523.13.3.384
- Martino, G., & Marks, L. E. (1999). Perceptual and linguistic interactions in speeded classification: Tests of the semantic coding hypothesis. *Perception*, *28*, 903–923.
- Maurer, D., & Mondloch, C. (2005). Neonatal synesthesia: A re-evaluation. In L. Robertson & N. Sagiv (Eds.), *Synesthesia: Perspectives from cognitive neuroscience* (pp. 193–213). New York, NY: Oxford University Press.
- Melara, R. D. (1989). Dimensional interactions between color and pitch. *Journal of Experimental Psychology: Human Perception and Performance*, *15*, 69–79.
- Melara, R. D., & Marks, L. E. (1990). Processes underlying dimensional interactions: Correspondences between linguistic and non-linguistic dimensions. *Memory & Cognition*, *18*, 477–495. doi:10.3758/BF03198481
- Melara, R. D., & O'Brien, T. P. (1987). Interaction between synesthetically corresponding dimensions. *Journal of Experimental Psychology: General*, *116*, 323–336.
- Meredith, M. A., & Stein, B. E. (1986). Visual, auditory, and somatosensory convergence on cells in superior colliculus result in multisensory integration. *Journal of Neurophysiology*, *56*, 640–662.
- Miller, J. (1982). Divided attention: Evidence for coactivation with redundant signals. *Cognitive Psychology*, *14*, 247–279. doi:10.1016/0010-0285(82)90010-X
- Miller, J. (1991). Channel interaction and the redundant-targets effect in bimodal divided attention. *Journal of Experimental Psychology: Human Perception and Performance*, *17*, 160–169. doi:10.1037/0096-1523.17.1.160
- Mondloch, C. J., & Maurer, D. (2004). Do small white balls squeak? Pitch–object correspondences in young children. *Cognitive, Affective, & Behavioral Neuroscience*, *4*, 133–136. doi:10.3758/CABN.4.2.133
- Murray, E. A., & Mishkin, M. (1985). Amygdalectomy impairs cross-modal association in monkeys. *Science*, *228*, 604–606.
- Ngo, M. K., & Spence, C. (2010). Auditory, tactile, and multisensory cues facilitate search for dynamic visual stimuli. *Attention, Perception, & Psychophysics*, *72*, 1654–1665. doi:10.3758/APP.72.6.1654
- Parise, C., & Spence, C. (2008). Synaesthetic congruency modulates the temporal ventriloquism effect. *Neuroscience Letters*, *442*, 257–261.
- Parise, C. V., & Spence, C. (2009). “When birds of a feather flock together”: Synesthetic correspondences modulate audiovisual integration in non-synesthetes. *PLoS One*, *4*, e5664.
- Patching, G. R., & Quinlan, P. T. (2002). Garner and congruence effects in the speeded classification of bimodal signals. *Journal of Experimental Psychology: Human Perception and Performance*, *28*, 755–775.
- Pratt, C. C. (1930). The spatial character of high and low tones. *Journal of Experimental Psychology*, *13*, 278–285.
- Ramachandran, V. S., & Hubbard, E. M. (2001). Synaesthesia: A window into perception, thought and language. *Journal of Consciousness Studies*, *8*, 30–34.
- Roffler, S. K., & Butler, R. A. (1968). Localization of tonal stimuli in the vertical plane. *Journal of the Acoustical Society of America*, *43*, 1260–1266.
- Rusconi, E., Kwan, B., Giordano, B., Umiltà, C., & Butterworth, B. (2006). Spatial representation of pitch height: The SMARC effect. *Cognition*, *99*, 113–129.
- Sadaghiani, S., Maier, J. X., & Noppeney, U. (2009). Natural, metaphoric, and linguistic auditory direction signals have distinct influences on visual motion processing. *Journal of Neuroscience*, *29*, 6490–6499.
- Spence, C. (2011). Crossmodal correspondences: A tutorial review. *Attention, Perception, & Psychophysics*, *73*, 971–995. doi:10.3758/s13414-010-0073-7
- Spence, C. (2012). Multisensory perception, cognition, and behavior: Evaluating the factors modulating multisensory integration. In B.

- E. Stein (Ed.), *The new handbook of multisensory processing* (pp. 241–264). Cambridge, MA: MIT Press.
- Spence, C., & Ngo, M. (2012). Does crossmodal attention or multisensory integration explain the crossmodal facilitation of masked visual target identification in the freezing effect? In B. E. Stein (Ed.), *The new handbook of multisensory processing* (pp. 345–358). Cambridge, MA: MIT Press.
- Spence, C., Ngo, M., Lee, J.-H., & Tan, H. (2010). Solving the correspondence problem in haptic/multisensory interface design. In M. H. Zadeh (Ed.), *Advances in haptics* (pp. 47–74). Vukovar, Croatia: In-Tech Publishers. Retrieved on March 24, 2012, from [www.sciyo.com/articles/show/title/solving-the-correspondence-problem-in-haptic-multisensory-interface-design](http://www.sciyo.com/articles/show/title/solving-the-correspondence-problem-in-haptic-multisensory-interface-design)
- Stein, B. E., & Meredith, M. A. (1993). *The merging of the senses*. Cambridge, MA: MIT Press.
- Stevens, S. S. (1957). On the psychophysical law. *Psychological Review*, *64*, 153–181. doi:10.1037/h0046162
- 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*, 679–690. doi:10.1093/cercor/bhk016
- Trimble, O. C. (1934). Localization of sound in the anterior posterior and vertical dimensions of auditory space. *British Journal of Psychology*, *24*, 320–334.
- Van der Burg, E., Cass, J., Olivers, C. N. L., Theeuwes, J., & Alais, D. (2010). Efficient visual search from synchronized auditory signals requires transient audiovisual events. *PLoS One*, *5*, e10664.
- Van der Burg, E., Olivers, C. N. L., 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*, 1053–1065. doi:10.1037/0096-1523.34.5.1053
- Van der Burg, E., Talsma, D., Olivers, C. N. L., Hickey, C., & Theeuwes, J. (2011). Early multisensory interactions affect the competition among multiple visual objects. *NeuroImage*, *55*, 1208–1218.
- Wolfe, J. M., Cave, K. R., & Franzel, S. L. (1989). Guided search: An alternative to the feature integration model for visual search. *Journal of Experimental Psychology: Human Perception and Performance*, *15*, 419–433. doi:10.1037/0096-1523.15.3.419