Synesthetic Associations Between Voice and Gestures in Preverbal Infants: Weak Effects and Methodological Concerns

A. Anikin[1*], M. Rudling[1], T. Persson[1], P. Gärdenfors[1]

[1] Lund University Cognitive Science, Lund, Sweden

***Correspondence:**
Andrey Anikin
andrey.anikin@lucs.lu.se

Keywords: preferential looking, audiovisual congruency, cross-modal matching, infant, synesthesia

Abstract

Adult humans spontaneously associate visual features, such as size and direction of movement, with phonetic properties like vowel quality and auditory pitch. A number of recent studies have claimed that looking time in preverbal infants reveals the same associations, which would indicate that some cross-modal correspondences are the result of perceptual biases. Here we tested 30 infants of age 7-13 months, who were exposed to pairs of audiovisual stimuli presented first sequentially and then side by side. The stimuli consisted of a visual object (computer-animated ball or filmed human hand) moving sinusoidally, vertically, or in a U-shape and accompanied by a sliding voice-like tone. Sequential presentation revealed no preference for either audiovisual synchrony or synesthetic congruency, while side-by-side presentation revealed a small preference for incongruent stimuli. The effect of congruency was similar for the animated ball and filmed human hand. These findings extend the results of previous research on pitch-motion synesthesia in preverbal infants, which used animations and sliding whistles, to more ecologically relevant stimuli such as voice and gestures. If infants and adults share the same preferences for non-arbitrary mappings between manual gestures and intonation, this could indicate that cross-modal correspondences facilitate language acquisition. On the other hand, a critical survey of the field revealed that previous studies of audiovisual cross-modal correspondences in infants suffer from replication failures due to poor robustness of the reported effect with respect to experimental stimuli and testing procedure. We therefore argue that the research on cross-modal correspondences in infants would profit from using alternative testing methods in addition to preferential looking and call for replication of previously reported congruency effects.

1 Introduction

The brain receives sensory information from different modalities and integrates it into unified mental representations of the surroundings. A critical step in this process is to be able to judge which parts of sensory input refer to the same object or event. This cross-modal binding can be accomplished using spatiotemporal matching: stimuli from different modalities that originate from approximately the same location and/or time are perceived as the same event (Spence, 2011). The detection of spatiotemporal synchrony takes place at an early, precortical stage of sensory processing in the superior colliculus (Holmes & Spence, 2005), and it is well-documented both in young children (Bahrick et al., 2004; Bremner, 2011; Spelke, 1979; Streri et al., 2013) and in non-human animals (reviewed in Ratcliff et al., 2016).

Many studies have recently focused on another - more arbitrary and synesthesia-like - type of cross-modal correspondences. Following Spence (2011), we refer to systematic associations between non-redundant stimulus attributes in different modalities as “synesthetic congruency”. For example, rounded/spiky shapes are associated with meaningless words like “baluma/takete” (Köhler, 1929) or “bouba/kiki” (Ramachandran and Hubbard, 2001). Similarly, auditory pitch is associated with height, spikiness, brightness, weight, and size of visual objects (Evans & Treisman, 2010; Parise & Spence, 2012). Exposure to culture-specific stimuli could influence some of these associations: for example, the mapping of colors to letters may be learned through early exposure to a particular toy alphabet (Witthoft, Winawer, & Eagleman, 2015). It is therefore important to test for their presence in preverbal infants and non-human animals.

Evidence of synesthesic congruency in animals is scarce, although there are several reports of prothetic cross-modal correspondences – that is, of matching stimuli in different modalities based on the amount rather than quality of sensory experience (Spence, 2011). For example, primates associate visual size with increasing auditory loudness (Ghazanfar & Maier, 2009; Maier et al., 2004). There is also some evidence of metathetic associations, which are based on sensory quality and are thus synesthetic in character, such as associations between visual luminance and auditory pitch in chimpanzees (Ludwig et al., 2011). In human infants, synesthetic congruency has been described extensively. Infants of 3-12 months of age have been reported to associate pitch with vertical position or movement (Dolscheid et al., 2012; Jeschonek et al., 2012; Walker et al., 2010), thickness (Dolscheid et al., 2012), and spikiness (Jeschonek et al., 2012). There is also evidence of synesthetic congruency between particular phonemes and object properties such as size (Peña et al., 2011) and shape (Ozturk et al., 2013) in preverbal infants.

These findings suggest that some synesthetic cross-modal correspondences predate language in ontogeny, and possibly also in phylogeny. Unfortunately, the results of infant studies are sometimes inconsistent, or the effects disappear after a slight modification of experimental stimuli and procedure. For example, the association between pitch and vertical movement in the study by Jeschonek et al. (2012) was found for dynamic displays, but not for static displays, and for anticipatory looking, but not when both stimuli were shown side by side. Lewkowicz & Minar (2014) report a failure to replicate the results of the study by Walker et al. (2010), which showed that infants associate pitch with vertical movement and shape, using a much larger sample of participants (see the response by Walker et al., 2014, who contend that this was not a close replication). The bouba-kiki effect, which is well documented in both adults and toddlers, failed to be reproduced in preverbal infants in three experiments by Fort et al. (2013).

Since testing procedures and stimuli are not identical across these studies, it is hard to compare them directly. However, the core method is always some variety of the preferential looking paradigm. The most common procedure is to show two videos sequentially, one with a matching and the other with a non-matching accompanying sound (e.g., Walker et al., 2010; Dolscheid et al., 2012; Lewkowicz & Minar, 2014). An alternative approach is to present two videos side by side and let the participant decide which one matches the sound (e.g., Spelke, 1979; Lewkowicz & Minar, 2014; Ghazanfar & Maier, 2009). In a third method, the participant is habituated to both videos, followed by a test trial with anticipatory looking or visual search for the matching video (Jeschonek et al., 2013; Spelke, 1979). All three methods have sometimes demonstrated synesthetic congruency, but sometimes also failed to do so (Table 1). Furthermore, occasionally different testing procedures produced incompatible results within one study and for the same stimuli (Bremner et al., 2011; Jeschonek et al., 2013). As a consequence, there are some debates concerning whether audiovisual synesthetic correspondences in preverbal infants are relatively weak or even absent, or whether they are present but difficult to detect (e.g., Lewkowicz & Minar, 2014 vs. Walker et al., 2014).

To cast additional light on this issue, we generated a wide variety of audiovisual stimuli and presented them to 30 infants of age 7-13 months, either sequentially or side by side. Given the lack of consistent results in previous studies, we deemed it important to include a positive control. A null result may be obtained either because infants lack the ability to make a particular discrimination or because the method fails to detect it. In contrast, if a study includes stimuli that the infants are known to discriminate, a lack of effect impugns the method itself. In our study this positive control was achieved by including an object that moved in a sinusoidal pattern and was accompanied by a tone with a matching or mismatching rate of frequency modulation. Matching these auditory and visual stimuli requires only the ability to detect synchrony across modalities, rather than synesthetic congruency. Since this ability is well-documented in infants (Bahrick et al., 2004; Streri et al., 2013) and non-human animals (Ratcliff et al., 2016) and does not require cortical processing (Holmes & Spence, 2005), we reasoned that infants must be able to match these control stimuli. Accordingly, failing to observe an effect for these stimuli would cast doubt upon the testing procedure.

Apart from this control, we included a vertically bouncing ball accompanied by a sliding tone, aiming to approximately reproduce the experimental stimuli previously used by Walker et al. (2010) and Lewkowicz & Minar (2014). In addition, we wanted to investigate whether associations between pitch and vertical movement might extend to other movements with a vertical component, such as a U-shaped vertical trajectory. If that is the case, synesthetic congruency between pitch and movement may be relevant to a broad range of ecological stimuli that the infants are exposed to, including speech prosody accompanied by manual gestures. Both the vertical and the U-shaped stimuli are qualitatively different from the sinusoidal displays, since the former are fully synchronized in both congruent and incongruent versions and can only be discriminated if infants non-arbitrarily map pitch shifts onto the direction of vertical movement.

To test whether natural human gestures might profit from synesthetic congruency by engaging matching intonation contours, we duplicated all animations with videos of a human hand describing similar patterns of movement. In particular, we speculated that the ability of infants to match the intonation of utterances with hand gestures might facilitate language acquisition (Corballis, 2003; McNeill, 2012; Arbib, 2012). Moreover, since infants are constantly exposed to speaking and gesticulating humans, these stimuli should increase the ecological validity compared to the previously tested animations. The authenticity of experimental stimuli was further enhanced by means of

Table 1 Studies of audiovisual congruency in infants that measured looking time

Publication	Research question	Age (months)	N	Method	Main findings as reported	Congruency effect
Spelke, 1979	Spatiotemporal correspondence (rate and synchrony)	3.5 - 5	48	Side by side (familiarization)	Singif. in two sessions, non-signif. in four sessions	Mixed
				Visual search	Signif. for first look, non-signif. for looking duration	Mixed
Walker et al., 2010	Pitch ~ vertical movement	4 - 5	16	Sequential	12/16 infants looked longer at congruent; looking time 28.5 ± 10.5 s for congruent and 21.7 ± 11.9 s for incongruent	Positive
	Pitch ~ spikiness	3 - 5	16	Sequential	12/16 infants looked longer at congruent; looking time 20.1 ± 7.9 s for congruent and 15.9 ± 6.3 s for incongruent	Positive
Peña et al., 2011	Phonemes ~ size	4	56	Side by side	First look and total looking time: /a/ and /o/ = large size, /i/ and /e/ = small size	Positive
Bremner et al., 2011	Spatiotemporal correspondence	2 - 8	36	Sequential	Non-signif. for all age groups	None
			36	Sequential after habituation to congruent	Marginal in first block, signif. in second block	Mixed
			36	Sequential after habituation to incongruent	Signif. for age 2 months, but not 5 and 8 months	Mixed
Dolscheid et al., 2012	Pitch ~ vertical movement	4 - 5	20	Sequential	14/20 infants looked longer at congruent; looking time 31.7 ± 11.4 s for congruent and 26.1 ± 13.3 s for incongruent	Positive
	Pitch ~ thickness	4 - 5	20	Sequential	13/20 infants looked longer at congruent; looking time 24.4 ± 11.8 s for congruent and 19.4 ± 11.5 s for incongruent	Positive
Haryu et al., 2012	Pitch ~ brightness	10	16	Sequential after habituation to congruent	Looking time 8.4 s for congruent and 13.1 s for incongruent	Negative
	Pitch ~ size	10	16		Looking time 12.1 s for congruent and 10.7 s for incongruent	None
Fort et al., 2013	Phonemes ~ spikiness	4 - 6	71	Side by side	Non-signif.	None
Jeschonek et al., 2013	Pitch ~ vertical movement	7 - 12	71	Side by side (familiarization)	Looking time 2.17 ± 0.44 s for congruent and 2.18 ± 0.42 for incongruent	None
				Anticipatory looking (dynamic test)	Looking time 1.56 ± 0.38 s for congruent and 1.37 ± 0.37 s for incongruent	Positive
	Pitch ~ vertical position			Anticipatory looking (static test)	Looking time 1.56 ± 0.47 s for congruent and 1.54 ± 0.39 s for incongruent	None
Ozturk et al., (2013)	Phonemes ~ spikiness	4	12	Sequential	10/12 infants looked longer at incongruent; looking time 13.0 ± 2.0 s for congruent and 16.8 ± 2.3 s for incongruent	Negative
Lewkowicz & Minar, 2014 (replication of Walker et al., 2010)	Pitch ~ vertical movement	4 - 12	188	Sequential	No difference in looking time for any age	None
	Pitch ~ spikiness	4 - 8	45	Sequential	No difference in looking time for any age	None
	Pitch ~ vertical movement	4 - 14	147	Side by side	No difference in looking time for any age	None

Fernández-Prieto et al., 2015	Pitch ~ size	4	18	Side by side	Non-signif.	None
		6	18		Signif.	Positive
Pietraszewski et al., 2017	Pitch and formants ~ size	3 - 4.5	32	Sequential	Signif. for both pitch and formants, but stronger for pitch	Negative
Walker et al., 2018	Pitch ~ vertical movement	Newborns (2 days old)	12	Sequential	Looking time 64 s for congruent and 42 s for incongruent, 10/12 infants looked longer at congruent	Positive
Anikin et al., 2018	Spatiotemporal correspondence, pitch ~ vertical movement	7 - 13	30	Sequential (familiarization)	Looking time 6.0 ± 6.6 s for congruent and 6.1 ± 6.4 s for incongruent	None
				Side by side	Looking time 3.4 ± 2.4 s for congruent and 3.8 ± 2.5 s for incongruent	Negative

generating a complex, voice-like harmonic sound with formants corresponding to a vowel /a/, rather than simply a pure tone used in previous studies (Dolscheid et al., 2012; Fernández-Prieto et al., 2015; Walker et al., 2010). We reasoned that this broad variety of ecologically relevant stimuli, as well as measuring looking time for both sequential and side-by-side presentation, would provide rich data on the ability of preverbal infants to detect audiovisual congruency.

2 Methods

2.1 Participants

Participants were recruited by sending a letter to the caregivers of every child of the required age, within a certain geographic area. We tested 32 infants aged 46 ± 8 weeks (M ± SD, range 7 to 13 months). Of these, one was excluded from the analysis due to equipment failure and another due to unwillingness to participate. No participant was reported having impaired hearing or vision. Testing took place in August, 2017. Ethical approval was sought from the regional ethical review board of Lund university, which concluded that the study did not require a formal evaluation by the ethics committee.

2.2 Stimuli

Each stimulus consisted of (1) a synthesized vowel /a/ with variable pitch and loudness and (2) a congruent or incongruent video of a moving object. We created five stimuli types (Fig. 1), in which the moving object was either an animated ball (“ball” stimuli) or the arm of a real person, which was filmed moving in synchrony with the sound (“hand” stimuli). We thus produced ten different stimuli pairs, where each pair included one congruent and one incongruent stimulus. The visual appearance of the ball occluding an underlying grid (Fig. 1), as well as the dynamically modulated loudness of the soundtrack, were a close replication of the method described by Walker et al. (2010).

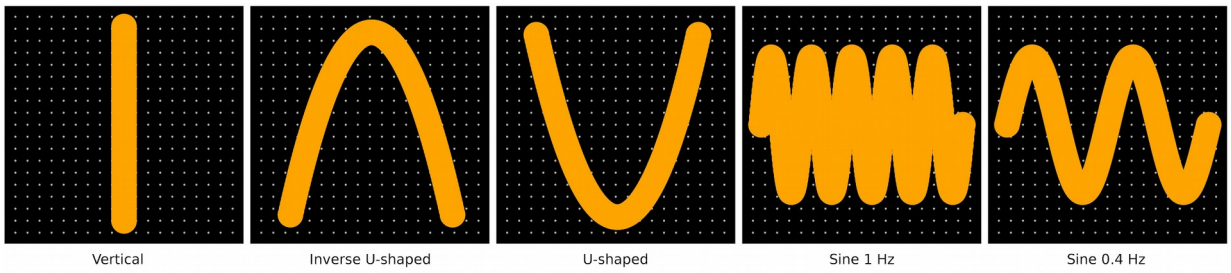

Figure 1. The patterns of movement followed by the visual object and mirrored by the pitch of the accompanying sound.

Stimuli of type “vertical” consisted of a visual object (ball or hand) moving up and down at a constant speed (Fig. 1), starting either at the top or the bottom of the screen. The sound always started at 400 Hz, decreased linearly on a logarithmic (musical) scale by one octave to 200 Hz over 2500 ms, and returned to 400 Hz over 2500 ms, with a 50 ms pause at each endpoint. The loudness varied linearly on a dB scale, reaching its maximum when the visual stimulus was in the middle and dropping by 26 dB when the object reached the top or bottom. The audiovisual display was considered congruent if the pitch increased as the object traveled upwards and incongruent otherwise.

Stimuli of the type “U-shaped” consisted of a visual object moving in a parabola, first left-to-right for 2500 ms and then right-to-left for 2500 ms, with a 50-ms pause at each endpoint. Horizontal speed was constant, while vertical speed varied quadratically, so the object appeared to slow down at the vertex and accelerate towards the endpoints. The soundtrack was similar to “vertical”, except that the pitch followed a parabolic, rather than a linear, trajectory on a logarithmic scale. The range of pitch varied from 200 Hz to 400 Hz. The loudness peaked at each end, where the perceived speed of movement was at its highest, and dropped by 26 dB at the vertex. In the congruent condition, both audio and video were of the same type (either “U” or “inverse U”); incongruent combinations were created by pairing up a “U” sound with an “inverse U” video, or vice versa.

For stimuli of the type “sine wave”, the visual object followed a sinusoidal trajectory, moving from left to right for 5000 ms and then right to left for 5000 ms, pausing for 50 ms at each endpoint. The pitch varied between 250 and 300 Hz, describing a sinusoidal trajectory on a logarithmic scale. The rate of frequency modulation (FM) was either 1 Hz or 0.4 Hz. The loudness varied sinusoidally at double the rate of FM, peaking as the pitch reached its mean value and dropping by 26 dB at the top and bottom of each FM cycle. Together with the sinusoidal variation of vertical speed, this reinforced the illusion of movement. In the congruent condition, both audio and video were of the same type (either “1 Hz” or “0.4 Hz”); incongruent combinations were created by pairing up “1 Hz” sound with “0.4 Hz” video, or vice versa.

All experimental sounds were created in R 3.0 (https://www.r-project.org) using a pre-release developmental version of *soundgen* package (Anikin, 2018). To achieve vowel-like quality, we synthesized a separate sine wave for each harmonic, with an exponential rolloff of harmonic power toward higher frequencies, and added four formants at 900, 1300, 2900, and 4300 Hz. The result approximated the vowel /a/ sung in a male voice. Ball animations were created in R by generating each frame separately. For the “hand” stimuli, a live human actor (A. A.) was filmed moving his arm in synchrony with the appropriate soundtrack against a uniform black background. Only the arm was visible in the movies (Fig. 2). The experiment was written in html/javascript and displayed in a web browser.

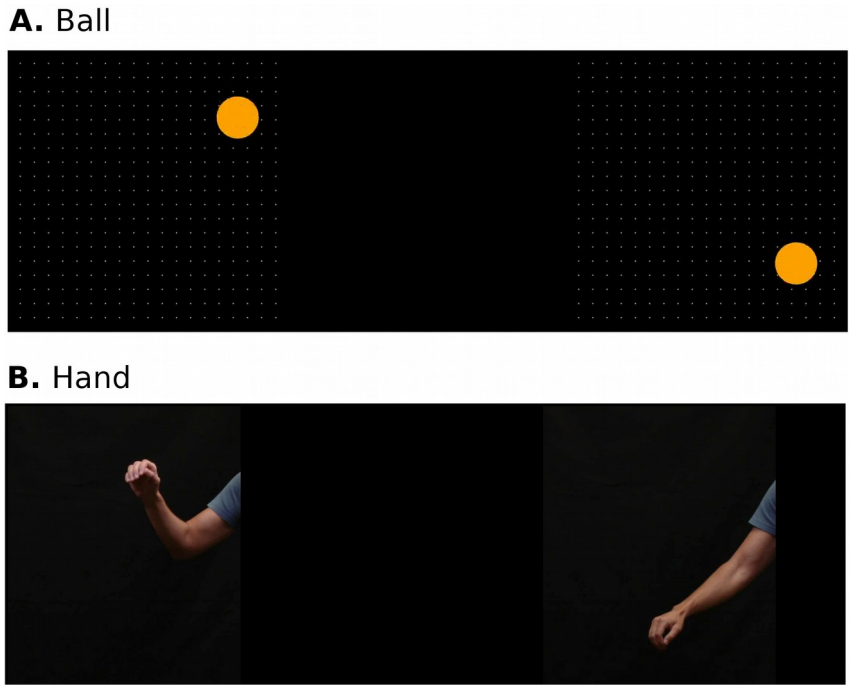

Figure 2. Two examples of what infants saw in the side-by-side phase, with “ball” (A) and “hand” (B) stimuli. We measured how long infants looked at the stimulus on the left vs. on the right.

2.3 Procedure

Written informed consent was given by a caregiver before testing commenced. Infants were seated in the lap of their parent, who was wearing ear muffs to avoid cueing the infant. Participants were facing a large screen, at 120 cm distance, while the experimenter sat to the infant's left and controlled the experiment by observing the infant's reactions and steering the protocol from her laptop. The order of trials, the order of presenting congruent and incongruent stimuli, and their left/right position on the screen were randomized for each new participant.

As shown in the flowchart in Figure 3, at the beginning of each trial an alert video (a pulsating star of shifting color accompanied by the sound of a baby laughing) was displayed until the experimenter ensured that the infant's attention was focused on the screen. Then two video clips with the same soundtrack were presented sequentially (phase 1); one of the videos was congruent with the soundtrack and the other incongruent. Each video was displayed until the infant averted their gaze for more than 1 s or showed clear inattention to the stimulus. The sound started and ended in synchrony with the videos. If the infant kept looking, the display continued for a maximum of 50 s, although very few trials were that long (only 13% of trials in phase 1 lasted over 10 s and 4% over 20 s). The maximum duration of 50 s was similar to that used in previous studies (60 s in Walker et al., 2010 and Bremner et al., 2011; 45 s in Haryu et al., 2012 and Pietraszewski et al., 2017; 40 s in Ozturk et al., 2013).

Following a brief interlude of 6 s, the fixation star was displayed again to attract the infant’s attention. Once the infant was looking at the screen, phase 2 began: the same two videos, with the same soundtrack as before, were now shown side by side, in the same location on the screen as in phase 1 (Fig. 2). They were displayed for 15 s, regardless of whether the infant was looking or not. A longer

intermission with another cartoon unrelated to the experiment concluded the trial. Each infant completed ten trials, one trial for each of ten stimuli pairs (5 stimuli types shown in Fig. 2 with a ball animation and 5 analogous types with a filmed hand), which took about 10 to 15 minutes. Two cameras recorded the procedure for manual frame-by-frame coding of gaze direction. The sound was muted during coding, so as to blind the coder (M. R.) to the experimental condition and thus avoid bias. A second person, blind to the experiment, independently coded 20% of the videos, demonstrating high inter-rater reliability (r = .96 for phase 1 and r = .91 for phase 2).

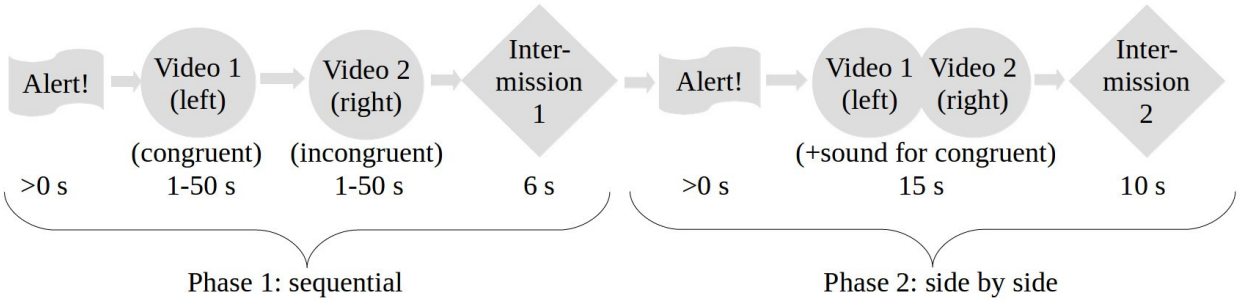

Figure 3. Flowchart of a single trial. Each infant completed ten such trials. The order of presentation was random when it came to both congruency and position on the screen.

2.4 Statistical analysis

All statistical tests and plotting were performed using R 3.4.3 (https://www.r-project.org). Preference for congruency seen as a binary outcome was analyzed with binomial tests and logistic mixed models, although the main analysis focused on looking time as a continuous variable. We used original, unaggregated trial-level data and fit mixed models to account for the nested nature of the data. Looking times in both sequential and side-by-side phases were strongly skewed, and they were therefore log-transformed prior to modeling. We added a small constant of 0.1 s to deal with zero values prior to log-transforming and then fit Gaussian mixed models with three random intercepts: per participant, per stimulus, and per trial (except when one of these was modeled as a fixed effect). Since looking time was on average several seconds (Fig. 4), and there were less than 4% of trials with zero looking time, adding 0.1 s had no significant effect on the results of any statistical analyses. Mixed models were fit using lme4 package (Bates, Maechler, Bolker, & Walker, 2015), and significance was tested using likelihood ratio tests.

We also replicated these mixed models and derived 95% credible intervals using Bayesian methods, which offer an in-built correction for multiple comparisons when simultaneously estimating the effect of congruency on looking time for ten different stimuli. This was achieved by specifying a horseshoe prior (Carvalho, Polson, & Scott, 2009) on regression coefficients, thus causing so-called shrinkage of regression coefficients towards zero and reducing the risk of false positives. Mildly informative conservative priors were used in models with a single fixed effect in order to improve the convergence of Markov Chain Monte Carlo sampling. A custom function was devised for log-transforming the axes in Figures 4 and 5 for both positive and negative values (for plotting purposes only). We created Bayesian models in Stan computational framework (http://mc-stan.org/) accessed with brms package (Buerkner, 2017).

3 Results

3.1 Sequential presentation

When a pair of congruent and incongruent audio-visual stimuli were presented sequentially (phase 1), infants looked longer at the congruent stimulus in 149 out of 300 trial pairs (30 participants x 10 trials). The observed preference for congruency was thus 49.7%, which is not significantly different from the chance level of 50% (binomial test: $p = .95$). Averaging per participant, the median proportion of trials in which infants looked longer at congruent stimuli was 50% (5 out of 10 trials), although there was considerable individual variability (from 2/10 to 9/10 trials).

Looking time was 6.0 ± 6.6 s (M ± SD) for congruent and 6.1 ± 6.4 s for incongruent stimuli (Fig. 4a). The most credible difference in the time infants looked at congruent vs. incongruent stimuli was 0.06 s, 95% CI [-0.68, 0.84], Cohen's *d* -0.002. Congruency thus had no overall effect on looking time (likelihood ratio test: $L = 0.02$, $df = 1$, $p = .88$), with no interaction between congruency and the type of stimulus (ball vs. hand: $L = 0.27$, $df = 1$, $p = .60$; sinusoidal vs. vertical vs. U-shaped: $L = 2.4$, $df = 5$, $p = .79$). Looking time decreased significantly over multiple trials ($L = 48.6$, $df = 1$, $p < .001$), and infants tended to look longer at whichever stimulus of a congruent-incongruent pair was presented first ($L = 20.2$, $df = 1$, $p < .001$). There was also a preference for looking longer at the stimulus on the right-hand vs. left-hand side of the screen when stimuli were shown one at a time ($L = 6.8$, $df = 1$, $p = .009$). Controlling for these confounds in a multiple regression model, congruency still had no effect on looking time ($L = 0.0001$, $df = 1$, $p = .99$). The age of infants had no effect on congruency preference (Congruency x Age interaction: $L = 0.03$, $df = 1$, $p = .86$).

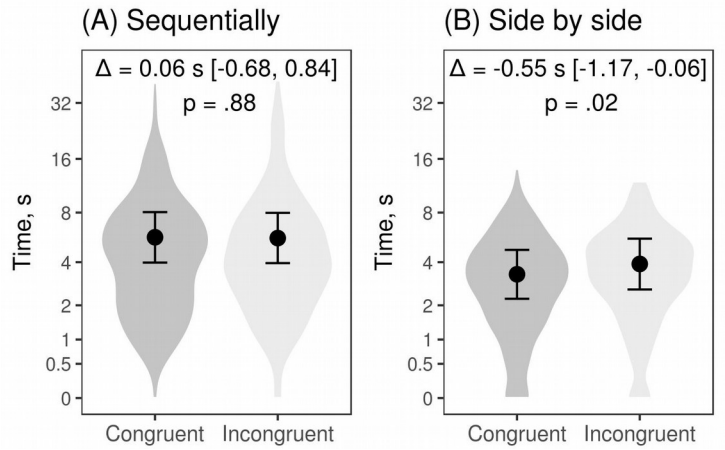

Figure 4. The time infants looked at congruent and incongruent audio-visual stimuli when these were presented sequentially (A) or side by side (B). Violin plots show the distribution of looking times. Solid points with error bars show fitted values from mixed models and 95% credible intervals.

There is thus no evidence that infants looked longer at congruent than at incongruent audio-visual combinations when these were presented sequentially. As shown in Figure 5a, this lack of congruency preference was observed both overall and for each stimuli type (Congruency x Stimulus interaction: $L = 2.4$, $df = 5$, $p = .78$). However, infants looked approximately 2 s longer at videos with the ball than with the hand ($L = 19.6$, $df = 1$, $p < .001$).

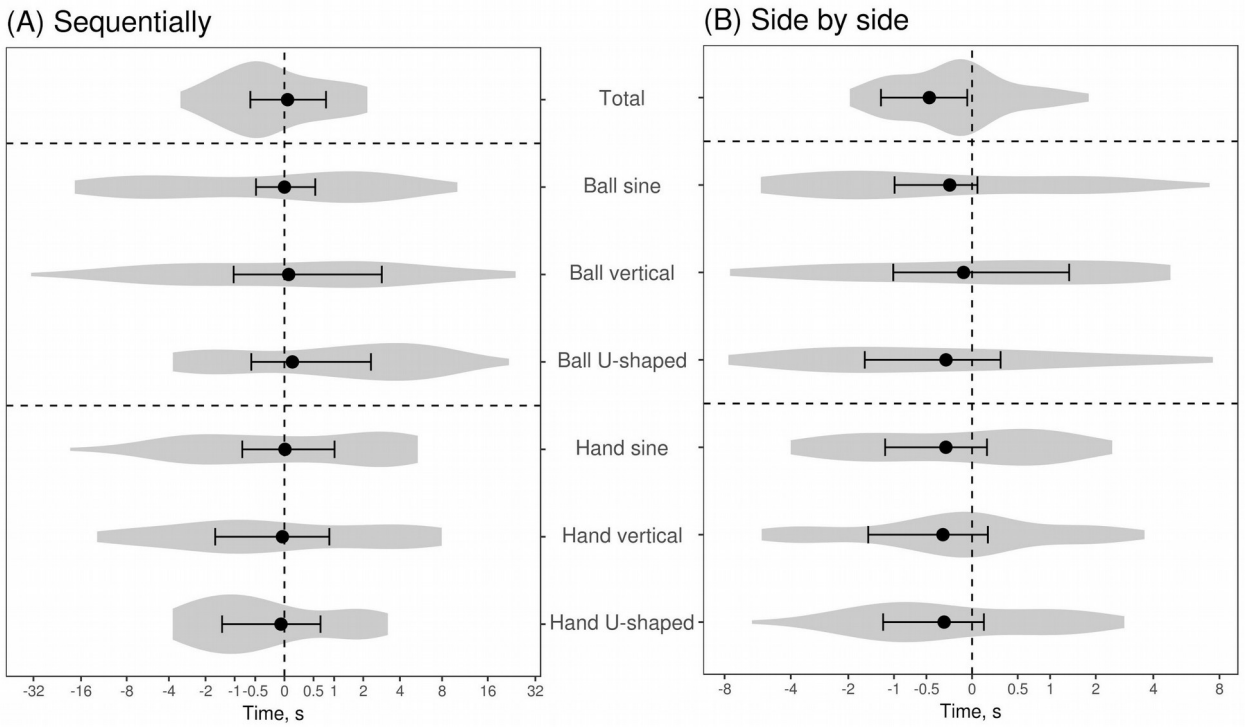

Figure 5. The difference in looking times for congruent vs. incongruent stimuli of different types, presented sequentially (A) and side by side (B). Violin plots show the distribution of observed values ("total" shows the distribution of median values per participant). Solid points with error bars show fitted values from mixed models and 95% credible intervals.

3.2 Side-by-side presentation

When a pair of congruent and incongruent audio-visual stimuli were presented side by side, infants looked longer at the congruent than at the incongruent stimuli in 143 out of 300 trials (47.7%, binomial test, $p = .45$). Averaging per participant, preference for congruent stimuli was again inconsistent (median 5 out of 10 trials, range from 2/10 to 10/10). Since the stimuli in each trial appeared on the same side of the screen first sequentially and then side-by-side, the infants had an opportunity to learn the location of the congruent stimulus in each pair. We therefore hypothesized that infants might manifest their preference for congruency by first looking in the direction of the congruent stimulus in the second, side-by-side phase of each trial. However, the congruent stimulus was the first to be looked at in only 139 out of 284 trials, or 48.9% (binomial test, $p = .77$; the number of trials is different from 300, since in some side-by-side trials the infant did not look at either video).

Infants spent on average 3.4 ± 2.4 s (M ± SD) looking at the congruent stimulus and 3.8 ± 2.5 s looking at the incongruent stimulus in each pair (Fig. 4b). The most credible difference in the time infants looked at congruent vs. incongruent stimuli was -0.55 s (95% CI [-1.17, -0.06]), Cohen's d -0.17. This suggests that infants displayed a small, but statistically significant preference for looking at incongruent stimuli ($L = 5.1$, $df = 1$, $p = .02$). This preference for incongruent stimuli did not depend on the stimulus type (ball vs. hand: $L = 0.21$, $df = 1$, $p = .64$; sinusoidal vs. vertical vs. U-shaped: $L = 2.2$, $df = 5$, $p = .82$; see Fig. 5b) or on the order in which the congruent and the incongruent stimuli pairs were presented in phase 1 (Congruency x Order interaction: $L = 1.0$, $df = 1$, $p = .31$). However, the effect of congruency was relatively small, and excluding three restless infants, who presented the

greatest challenges with coding, weakened it further (L = 4.0, df = 1, p = .05). The age of infants had no effect on congruency preference (Congruency x Age interaction: L = 0.11, df = 1, p = .74).

As with sequential presentation, looking time decreased over progressive trials (L = 23.3, df = 1, p < .001), but this time there was a slight preference for the left-hand side of the screen (L = 6.7, df = 1, p = .01). Controlling for trial number and position on the screen, congruency still had an effect on looking time (L = 4.9, df = 1, p = .03). As with sequential presentation, infants looked longer at the stimuli with a ball than with a hand (L = 11.6, df = 1, p < .001).

Based on preferential looking in a side-by-side setting, there is thus some evidence that infants looked slightly longer at incongruent stimuli.

4 Discussion

Informed by previous research, we tested the hypothesis that preverbal infants would reveal their sensitivity to audiovisual congruency through differences in looking time between congruent and incongruent displays. Sequential presentation of one congruent and one incongruent display failed to reveal an effect of congruency for any of the tested stimulus types, including an object that was moving sinusoidally (audiovisual matching based on spatiotemporal synchrony) or bouncing up and down vertically or in a parabola (synesthetic mapping of auditory pitch to vertical displacement). When the same stimuli were then presented side by side following this familiarization with both congruent and incongruent displays, looking times were slightly longer for incongruent stimuli. This effect was consistent across stimulus types, suggesting that incongruent audiovisual combinations may have attracted the infants' attention by virtue of violating their expectations.

This study had two novel features: improved ecological validity of experimental stimuli and a new two-step testing procedure. To begin with the stimuli, we used a synthetic, but highly realistic, voice-like sound source. It is perceptually distinct from sliding whistles, which are much easier to synthesize and therefore typically paired with dynamic visual displays (e.g., Dolscheid et al., 2012; Fernández-Prieto et al., 2015; Walker et al., 2010). The synthetic voice we used was closer in quality to actual recordings of human speakers, which are normally employed when precise temporal modulation of the sound is not required - for example, in bouba-kiki studies (e.g., Ozturk et al., 2013). The intonation of a synthetic voice can be controlled as precisely as that of synthetic whistles, but without compromising the vowel-like quality. Flexible and easily programmed voice synthesizers such as *soundgen* (Anikin, 2018) are thus potentially useful for researchers studying audiovisual associations, since they combine a more voice-like harmonic sound with standardization and perfect control needed for rigorous testing.

As for the visual stimuli, we tested both animations of a bouncing ball, as in earlier studies (e.g., Dolscheid et al., 2012; Walker et al., 2010), and a previously untested visual stimulus, namely filmed movements of a person's arm. An important limitation of using a live actor is the difficulty of achieving sufficiently high synchrony between manual gestures and the accompanying soundtrack. A combination of computer animation with a voice synthesizer may ultimately prove a better method of generating ecologically relevant audiovisual stimuli. Nevertheless, we found no evidence that the infants were affected by the potentially imperfect synchronization between the actor's movements and the soundtrack, since the effects of congruency were similar for the ball and hand videos. On the other hand, the infants looked longer at the ball than at the arm in both the congruent and the incongruent condition. Overall, our findings extend the results of previous studies, which predominantly used

artificial stimuli like sliding whistles and computer animations, to real-world audiovisual displays such as human voice coupled with manual gestures.

The second novel feature of our study was its testing procedure, in which infants first watched a congruent and an incongruent audiovisual display sequentially and then viewed both side by side. Looking behavior has previously been evaluated after habituation to congruent (Haryu, 2012), incongruent (Bremner, 2011), or both stimuli pairs (Spelke, 1979; Jeschonek et al., 2013), and our procedure can be seen as a new variation of this technique. The first phase of each trial was similar to the procedure employed in the influential study by Walker et al. (2010), and simultaneously it served as a familiarization phase followed by preferential looking. We were particularly interested in comparing the results from these two phases. Walker and coauthors (2010) reported longer looking times for congruent stimuli, but this effect could not be replicated in a large study by Lewkowicz and Minar (2014). We also observed no preference for congruent stimuli in the sequential phase. The average looking time (~6 s) was shorter than previously reported, which may have been due to differences in the criteria for terminating a trial and in the ability of the tested audiovisual displays to hold the infants' attention. However, this should not have prevented us from detecting a prefence for congruent displays, which was completely absent for all stimuli types. Notably, an effect of congruency was absent for the sinusoidal stimuli, which we treated as positive controls and which could be matched purely based on audiovisual synchrony rather than synesthetic correspondences - an ability that is well-established in both infants (Bahrick et al., 2004; Holmes & Spence, 2005; Streri et al., 2013) and animals (Ratcliff et al., 2016). The lack of congruency effects for the sinusoidal stimuli strongly suggests that the method of sequential presentation with no prior habituation failed to reveal existing preferences. In contrast, the second phase in each trial, in which the same stimuli were then presented side by side, revealed a small, but significant preference for incongruent stimuli.

A preference for looking at incongruent stimuli has been described in infants with both side-by-side (Ozturk et al., 2013) and sequential (Haryu et al., 2012) presentation. However, there are also multiple reports of infants looking longer at congruent stimuli (e.g., Dolscheid et al., 2012; Peña et al., 2011). Taken at face value, our results indicate that familiarization with both stimuli during sequential presentation may then encourage looking at the mismatching stimulus in side-by-side presentation. However, the effect size was very small (Cohen’s d = -0.17) for all types of stimuli, including the sinusoidal control. Assuming that infants can detect spatiotemporal synchrony needed to match the sinusoidal stimuli, the experimental method was not very sensitive. In addition, it is not entirely clear how familiarization in phase 1 may have influenced looking behavior in phase 2. The outcome may have been different if two independent samples of infants had been tested instead of this two-step procedure. Considering these methodological limitations of the present study, we hesitate to interpret our findings as unambiguous evidence that infants detect audiovisual congruency in the tested stimuli, revealing this ability by looking longer at incongruent displays. Instead, we argue that the present results are best seen as the latest contribution to a growing list of mixed and complicated findings (Table 1), which collectively question the robustness of preferential looking as a measure of cross-modal integration in human infants.

There are at least two possible explanations for these inconsistencies. The first is that cross-modal correspondences are learned and develop late, at a verbal stage (Lewkowicz & Minar, 2014). If this is true, the reported positive results arise due to chance, biased testing procedures, or incorrect data analysis. A second possibility is that preverbal infants are indeed sensitive to cross-modal correspondences, but this capacity is easily missed due to insufficient robustness of the testing

method or lack of statistical power. Null results were predominantly reported in the studies with the largest samples (Table 1), notably Lewkowicz & Minar (2014) and Fort et al. (2013), indicating that statistical power is unlikely to be the issue. A more likely explanation is that the testing method is not particularly robust, in the sense that the outcome is highly dependent on the testing procedure and stimuli, while the rate of false negatives is so high that a null result is not informative. Discussions of contradictory results have focused on subtle characteristics of experimental stimuli (Walker et al., 2014 vs. Lewkowicz, 2014), the sequence in which the stimuli were presented (Jeschonek et al., 2013), or the overall complexity of experimental design (Fort et al., 2013; Jeschonek et al., 2013; Ozturk et al., 2013). It is interesting to investigate the effect of all these factors on looking behavior. Nevertheless, the unpredictability of preferential looking is problematic for research when it is taken as a *prima facie* standard method for measuring infant discrimination.

This unpredictability is also the reason why many studies, including this one, employ multiple testing procedures with the same stimuli, often with conflicting results (Table 1), even though it requires larger samples or complicates the interpretation of results if the same infants perform multiple tests. Since the rate of false positives increases with multiple comparisons, the results of multi-stage tests - with the same or different samples - are also less reliable statistically. A related cause for concern is that the difference in looking time can be observed in either direction. If infants look longer at the congruent stimuli, the result is interpreted as preference for congruency (“positive” effect in Table 1), and if they look longer at the incongruent stimulus, the effect is couched in terms of violation of expectations (“negative” effect in Table 1). There are some theoretical considerations as to which of these two effects can be expected in particular circumstances. When tested repeatedly on a task, infants may initially look longer at familiar stimuli, while repeated exposure to the same stimuli shifts the preference to novel stimuli (Houston-Price & Nakai, 2004; Roder et al., 2000). Assuming that congruent stimuli are more familiar than incongruent stimuli, since they better correspond to the infant’s prior sensory experience and expectations, habituation may increase the preference for incongruent stimuli, possibly explaining why infants in our study looked longer at incongruent stimuli in phase 2, but not in phase 1. However, the relative contribution of familiarity and novelty effects, and thus the preference for congruent or incongruent stimuli, may also depend on the complexity of experimental stimuli and on the task. Complex stimuli and side-by-side presentation may favor congruent multimodal displays, while simple stimuli and sequential presentation may favor novel, incongruent combinations (Ozturk et al., 2013).

Considering how many factors may affect the balance between familiarity and novelty effects, in practice it is difficult to predict the direction of congruency effect in each particular experimental setting prior to data collection. An often overlooked consequence of this fact is the inadvisability of reporting the significance of differences in looking times using one-tailed tests (e.g., binomial tests in Walker et al., 2010). A two-tailed criterion should be used instead, considering the intrinsic difficulties of predicting the direction of the effect. We are not questioning all evidence of cross-modal congruency detection based on preferential looking in infants. For example, looking times in the same study by Walker et al. (2010) show a clear effect of congruency. Given the number of failed replications and weak or inconsistent findings, however, simply collecting more evidence of the same sort may no longer be enough: the key published results should be corroborated by evidence obtained with alternative methods. In adults, cross-modal associations have been investigated using a variety of techniques, including the implicit associations test (Parise & Spence, 2012), event-related potentials (Asano et al., 2015), and speeded classification (Jamal et al., 2017). Animal studies of cross-modal correspondences have likewise adopted a range of methods, such as match-to-sample (Izumi &

Kojima, 2004) and speeded classification (Ludwig et al., 2011), in addition to preferential looking (Evans et al., 2005; Maier, 2004). A similar diversification of methods would be welcome in research on infant cross-modal perception, and several alternative measures have already been tested in infants. For example, Nelson et al. (1993) employed EEG to detect haptic-visual matching, while Lewkowicz & Turkewicz (1980) proved that infants match the intensity of light and noise using cardiac response. It would be very useful to replicate the research on audiovisual synesthetic associations in infants using some of these alternative methods.

To summarize, two measures of looking time that we utilized in this study have provided conflicting evidence on the ability of preverbal infants to detect audiovisual congruency. Like Lewkowicz & Minar (2014), we failed to replicate an association between pitch and vertical movement using sequential presentation as reported by Walker et al. (2010) and Dolscheid et al. (2012). Presenting the same stimuli side by side after familiarization, however, did reveal a small preference for incongruent stimuli. These results indicate that preverbal infants may indeed possess synesthetic associations between sound frequency and the movements of inanimate objects, as previously observed (Table 1). In addition, we have demonstrated that such associations may also hold for ecologically valid stimuli, namely the intonation of a human voice and manual gestures. If further confirmed, these findings are of importance for those theories of language acquisition which propose that non-arbitrary mappings between intonation and gestures could facilitate learning new concepts (Corballis, 2003; McNeill, 2012; Arbib, 2012). However, the effect size was small, and a critical survey of the field reveals that audiovisual synesthetic associations in infants are detected less reliably than a casual review of the literature would suggest. We therefore caution researchers against relying on preferential looking as the predominant paradigm in studies of cross-modal correspondences in infants. Acknowledging the central importance of this method for research on the cognitive abilities of preverbal infants (Golinkoff et al., 2013) and non-human animals (Ratcliff et al., 2016), it is still desirable to substantiate - and perhaps to re-evaluate - the claims that preverbal infants map auditory pitch onto vertical movement and object properties such as shape and size.

5 Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

6 Author contributions

P.G. conceived the experiment. A.A, M.R, T.P., and P.G. designed the experiment. M.R. collected the data. A.A. analyzed the data and drafted the manuscript. A.A, M.R, T.P., and P.G. approved the final version of the manuscript.

7 Funding

The authors gratefully acknowledge support from the Swedish Research Council in the Linnaeus environment "Thinking in Time: Cognition, Communication and Learning". The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. We are also grateful to Marianne Guldberg for helping with study design and to Lund University Humanities Laboratory.

8 Data availability statement

The stimuli, the scripts used to produce them, and the dataset for this study can be downloaded from http://cogsci.se/publications.html.

9 References

Anikin, A. (2018). Soundgen: An open-source tool for synthesizing nonverbal vocalizations. *Behavior Research Methods*, 1-15.

Arbib, M. (2012). *How the brain got language: The mirror system hypothesis*. Oxford: Oxford University Press.

Asano, M., Imai, M., Kita, S., Kitajo, K., Okada, H., & Thierry, G. (2015). Sound symbolism scaffolds language development in preverbal infants. *Cortex*, *63*, 196-205.

Bahrick, L. E., Lickliter, R., & Flom, R. (2004). Intersensory redundancy guides the development of selective attention, perception, and cognition in infancy. *Current Directions in Psychological Science*, *13*(3), 99-102.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1-48.

Bremner, J. G., Slater, A. M., Johnson, S. P., Mason, U. C., Spring, J., & Bremner, M. E. (2011). Two to Eight Month Old Infants' Perception of Dynamic Auditory–Visual Spatial Colocation. *Child Development*, *82*(4), 1210-1223.

Bürkner, P. C. (2016). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*(1), 1-28.

Carvalho, C. M., Polson, N. G., & Scott, J. G. (2009, April). Handling sparsity via the horseshoe. In *Artificial Intelligence and Statistics* (pp. 73-80).

Corballis, M. (2003). *From hand to mouth: The origins of language*. Princeton: Princeton University Press.

Dolscheid, S., Hunnius, S., Casasanto, D., & Majid, A. (2012). The sound of thickness: Prelinguistic infants' associations of space and pitch. In: Miyake, N., Peebles, D., & Coope, R. (eds.), *Proceedings of the 34th Annual Conference of the Cognitive Science Society*; pp. 306-311.

Evans, K. K, & Treisman, A. (2009). Natural cross-modal mappings between visual and auditory features. *Journal of vision*, *10*(1):6, 1-12.

Evans, T. A., Howell, S., & Westergaard, G. C. (2005). Auditory-visual cross-modal perception of communicative stimuli in tufted capuchin monkeys (Cebus apella). *Journal of Experimental Psychology: Animal Behavior Processes*, *31*(4), 399-406.

Fernández-Prieto, I., Navarra, J., & Pons, F. (2015). How big is this sound? Crossmodal association between pitch and size in infants. *Infant Behavior and Development*, *38*, 77-81.

Fort, M., Weiß, A., Martin, A., & Peperkamp, S. (2013). Looking for the bouba-kiki effect in prelexical infants. In *Auditory-Visual Speech Processing (AVSP) 2013*.

Ghazanfar, A. A., & Maier, J. X. (2009). Rhesus monkeys (Macaca mulatta) hear rising frequency sounds as looming. *Behavioral Neuroscience*, *123*(4), 822-827.

Golinkoff, R. M., Ma, W., Song, L., & Hirsh-Pasek, K. (2013). Twenty-five years using the intermodal preferential looking paradigm to study language acquisition: What have we learned?. *Perspectives on Psychological Science*, *8*(3), 316-339.

Haryu, E., & Kajikawa, S. (2012). Are higher-frequency sounds brighter in color and smaller in size? Auditory–visual correspondences in 10-month-old infants. *Infant Behavior and Development*, *35*(4), 727-732.

Holmes, N. P., & Spence, C. (2005). Multisensory integration: space, time and superadditivity. *Current Biology*, *15*(18), R762-R764.

Houston Price, C., & Nakai, S. (2004). Distinguishing novelty and familiarity effects in infant preference procedures. *Infant and Child Development*, *13*(4), 341-348.

Izumi, A., & Kojima, S. (2004). Matching vocalizations to vocalizing faces in a chimpanzee (Pan troglodytes). *Animal Cognition*, *7*(3), 179-184.

Jamal, Y., Lacey, S., Nygaard, L., & Sathian, K. (2017). Interactions Between Auditory Elevation, Auditory Pitch and Visual Elevation During Multisensory Perception. *Multisensory Research*, *30*(3-5), 287-306.

Jeschonek, S., Pauen, S., & Babocsai, L. (2013). Cross-modal mapping of visual and acoustic displays in infants: The effect of dynamic and static components. *European Journal of Developmental Psychology*, *10*(3), 337-358.

Köhler, W. (1929). *Gestalt Psychology*. New York: Liveright.

Lewkowicz, D. J., & Minar, N. J. (2014). Infants are not sensitive to synesthetic cross-modality correspondences: A comment on Walker et al.(2010). *Psychological Science*, *25*(3), 832-834.

Lewkowicz, D. J., & Turkewitz, G. (1980). Cross-modal equivalence in early infancy: Auditory–visual intensity matching. *Developmental Psychology*, *16*(6), 597-607.

Ludwig, V. U., Adachi, I., & Matsuzawa, 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*(51), 20661-20665.

Maier, J. X., Neuhoff, J. G., Logothetis, N. K., & Ghazanfar, A. A. (2004). Multisensory integration of looming signals by rhesus monkeys. *Neuron*, *43*(2), 177-181.

McNeill, D. (2012). *How language began: Gesture and speech in human evolution*. Cambridge: Cambridge University Press.

Nelson, C. A., Henschel, M., & Collins, P. F. (1993). Neural correlates of cross-modal recognition memory by 8-month-old human infants. *Developmental Psychology*, *29*(3), 411-420.

Ozturk, O., Krehm, M., & Vouloumanos, A. (2013). Sound symbolism in infancy: evidence for sound–shape cross-modal correspondences in 4-month-olds. *Journal of Experimental Child Psychology*, *114*(2), 173-186.

Parise, C. V., & Spence, C. (2012). Audiovisual crossmodal correspondences and sound symbolism: a study using the implicit association test. *Experimental Brain Research*, *220*(3-4), 319-333.

Peña, M., Mehler, J., & Nespor, M. (2011). The role of audiovisual processing in early conceptual development. *Psychological Science*, *22*(11), 1419-1421.

Pietraszewski, D., Wertz, A. E., Bryant, G. A., & Wynn, K. (2017). Three-month-old human infants use vocal cues of body size. In *Proc. R. Soc. B*, *284*: 20170656.

Ramachandran, V. S., & Hubbard, E. M. (2001). Synaesthesia – a window into perception, thought and language. *Journal of Consciousness Studies*, *8*(12), 3-34.

Ratcliffe, V. F., Taylor, A. M., & Reby, D. (2016). Cross-modal correspondences in non-human mammal communication. *Multisensory Research*, *29*(1-3), 49-91.

Roder, B. J., Bushnell, E. W., & Sasseville, A. M. (2000). Infants' preferences for familiarity and novelty during the course of visual processing. *Infancy*, *1*(4), 491-507.

Spelke, E. S. (1979). Perceiving bimodally specified events in infancy. *Developmental Psychology*, *15*(6), 626-636.

Spence, C. (2011). Crossmodal correspondences: A tutorial review. *Attention, Perception, & Psychophysics*, *73*(4), 971-995.

Streri, A., de Hevia, M. D. D., Izard, V., & Coubart, A. (2013). What do we know about neonatal cognition?. *Behavioral Sciences*, *3*(1), 154-169.

Walker, P., Bremner, J. G., Mason, U., Spring, J., Mattock, K., Slater, A., & Johnson, S. P. (2014). Preverbal infants are sensitive to cross-sensory correspondences: much ado about the null results of Lewkowicz and Minar (2014). *Psychological Science*, *25*(3), 835-836.

Walker, P., Bremner, J. G., Mason, U., Spring, J., Mattock, K., Slater, A., & Johnson, S. P. (2010). Preverbal infants' sensitivity to synaesthetic cross-modality correspondences. *Psychological Science*, *21*(1), 21-25.

Walker, P., Bremner, J. G., Lunghi, M., Dolscheid, S., D. Barba, B., & Simion, F. (2018). Newborns are sensitive to the correspondence between auditory pitch and visuospatial elevation. *Developmental Psychobiology*, *60*(2), 216-223.

Witthoft, N., Winawer, J., & Eagleman, D. M. (2015). Prevalence of learned grapheme-color pairings in a large online sample of synesthetes. *PLoS One*, *10*(3), e0118996.