Bimodal emotion congruency is critical to preverbal infants’ abstract rule learning

Angeline Sin Mei Tsui,¹,² Yuen Ki Ma,¹ Anna Ho,¹ Hiu Mei Chow¹,³ and Chia-huei Tseng¹

¹. Department of Psychology, The University of Hong Kong, China,
². Department of Psychology, University of Ottawa, Canada,
³. Department of Psychology, University of Massachusetts Boston, USA

Abstract

Extracting general rules from specific examples is important, as we must face the same challenge displayed in various formats. Previous studies have found that bimodal presentation of grammar-like rules (e.g. ABA) enhanced 5-month-olds’ capacity to acquire a rule that infants failed to learn when the rule was presented with visual presentation of the shapes alone (circle-triangle-circle) or auditory presentation of the syllables (la-ba-la) alone. However, the mechanisms and constraints for this bimodal learning facilitation are still unknown. In this study, we used audio-visual relation congruency between bimodal stimulation to disentangle possible facilitation sources. We exposed 8- to 10-month-old infants to an AAB sequence consisting of visual faces with affective expressions and/or auditory voices conveying emotions. Our results showed that infants were able to distinguish the learned AAB rule from other novel rules under bimodal stimulation when the affects in audio and visual stimuli were congruently paired (Experiments 1A and 2A). Infants failed to acquire the same rule when audio-visual stimuli were incongruently matched (Experiment 2B) and when only the visual (Experiment 1B) or the audio (Experiment 1C) stimuli were presented. Our results highlight that bimodal facilitation in infant rule learning is not only dependent on better statistical probability and redundant sensory information, but also the relational congruency of audio-visual information. A video abstract of this article can be viewed at: https://m.youtube.com/watch?v=KYTyjH1k9RQ

Research highlights

- Eight- to ten-month-old infants can extract sequence rules (e.g. A-A-B) presented by audio-visual emotional cartoon faces, but not when the audio or visual sequence is presented alone.
- Preverbal infants possess the capacity to integrate audio-visual emotions at our tested age (8–10 months old).
- The congruency of audio-visual emotions in a cartoon face enhances infants’ benefits from bimodal learning.

Introduction

We receive versatile and continuous sensory stimulation from our environments, mostly conveyed by multiple sensory modalities. To efficiently interact with our environment, we need to extract the invariant information (e.g. an approaching car) contained in various forms (enlarging visual retinal image of an approaching car and looming sound from the engine) and apply it to new combinations of surroundings (e.g. an approaching truck at another intersection). This ability is essential to our daily lives, and even preverbal infants must possess it to survive.

Infants are able to extract and generalize simple grammatical rules from sensory stimuli. In a study conducted by Marcus, Vijayan, Bandi Rao and Vishton (1999), 7-month-old infants were first habituated to three-syllable sequence strings of AAB (e.g. ga ga ti), ABB (e.g. ga ti ti), or ABA (e.g. ga ti ga) rules. Later in the test, infants were able to distinguish the novel syllable strings that were either consistent (e.g. na li li) or
inconsistent (e.g. na li na) with the habituated structure (e.g. ABB). All syllables were novel in the test so that infants’ successful discrimination of a new rule was not a reflection of memory of exact items but a true acquisition of the target abstract rule. Studies have shown that infants at 7 to 8 months can also learn these rules from nonlinguistic stimuli – such as animal pictures, music tones, animal sounds, and geometric shapes (Saffran, Pollak, Seibel & Shkolnik, 2007; Marcus, Fernandes & Johnson, 2007; Fiser & Aslin, 2002; Johnson, Fernandas, Frank, Kirkham, Marcus et al., 2009; Kirkham, Slemmer & Johnson, 2002) – suggesting that infants’ ability to extract abstract rules is not specific to linguistic or audio stimuli but more likely a general ability operating across sensory domains domain-general ability across senses.

However, not all rules and stimulus types are equally accessible to preverbal infants during their development. In terms of stimuli, some studies have suggested that young infants might more readily learn rules from linguistic stimuli than nonlinguistic stimuli, possibly due to infants’ high sensitivity to linguistic components (Marcus, Johnson, Fernandes & Slemmer, 2004; Marcus et al., 2007). Within the visual modality, infants learn a new rule more easily when it is presented with familiar images (e.g. pictures of different dog breeds) than unfamiliar images (e.g. visual shapes) (Marcus et al., 2004; Saffran et al., 2007). Among all the abovementioned rules, AAB appears to be more difficult than ABA or ABB, possibly because it contains an early repetition in the sequence and demands more cognitive processing resources (Johnson et al., 2009). Behaviorally, 8-month-old infants are able to extract ABA and ABB rules – but not the AAB rule – from shape sequences; infants need to be at least 11 months old to learn the AAB rule. These variations reflect not only the complex nature of abstract rule acquisition for preverbal infants, but also the opportunities available to enhance the learning process.

While the above evidence is derived from unisensory domains, relatively few studies have examined how multisensory stimuli might affect infant learning. Real-world structure and rules are mostly represented in more than one sensory modality; it is therefore critical to investigate multimodal situations to identify optimal learning environments for young infants. Frank, Slemmer, Marcus and Johnson (2009) showed that presenting syllables synchronously with visual shape sequences facilitates 5-month-olds’ acquisition of ABA and ABB rules that they were unable to acquire in unisensory domains. Likewise, a more recent study by Thiessen (2012) also demonstrated the power of bimodal stimulation (tone and shape) in infant rule learning from non-linguistic stimuli.

Visual and audio temporal synchrony has been identified to be critical to facilitate infants’ perceptual detection and recognition (Gogate & Bahrick, 1998; Bahrick, Flom & Lickliter, 2002; Bahrick, Lickliter & Flom, 2004; Lewkowicz, 2000; Flom & Bahrick, 2007; see Bahrick & Lickliter, 2012, for a review). For example, when a toy hammer taps a surface with synchronous sights and sounds, infants detect the rhythm and tempo accurately, but they do not when it is presented in visual or auditory modalities alone, or when the sights and sounds are out of synchrony (Bahrick & Lickliter, 2002). Evidence also shows that 4-month-old infants are able to discriminate the serial order of moving objects generating visible and audible impacts only when the event is specified multimodally (Lewkowicz, 2004). One prominent theory, the Intersensory Redundancy Hypothesis (IRH) proposes that information presented concurrently and synchronously across multiple modalities receives infants’ attentional priority, thus resulting in better recognition (Bahrick, 2010; Bahrick & Lickliter, 2000, 2002; Bahrick et al., 2004; Lickliter & Bahrick, 2004). Hence, bimodal facilitation can be a result of our enhanced attention driven by audio-visual synchrony.

Similar bimodal facilitation has been reported in adult learning tasks (Seitz, Kim & Shams, 2006; Kim, Seitz & Shams, 2008). For example, Seitz, Kim & Shams (2006) trained observers to detect motion from two sequential visual displays. One visual display contained dots moving in random directions (noise), and the other had a small portion of dots moving in one direction (coherent motion) in addition to other randomly moving dots (noise). Observers reported which of the two visual displays contained the motion signal, and their performance steadily improved (measured by the minimum motion signal strength needed for an accurate report) in the course of five days of practice. When there was a sound present temporally synchronously with visual training (either as stationary noise or moving noise), participants’ learning within and across sessions was significantly better than with the purely visual presentation. This enhancement of visual motion sensitivity by simultaneous presentation of auditory moving stimuli is observed even when the auditory stimuli are task-irrelevant and present in both visual displays thus not informing the correct interval that contained the motion signal (Kim, Peters & Shams, 2012).

These results suggest a few possible origins of bimodal benefit from perceptual encoding, attentional enhancement and higher level reorganization of knowledge representation. When redundant information is carried by more than one sensory modality, it is not surprising that it has a greater statistical likelihood of being successfully encoded (Alais & Burr, 2004; Wuerger, Hofbauer & Meyer, 2003). This is similar to the analogy...
that we have a higher chance of finding an item from two shops (modalities) than from one shop only. Extra-sensory stimulation can also increase the alerting attentional system (Seitz et al., 2006) or alter the selective attentional system (Bahrick & Lickliter, 2012) because multiple sensory sources arouse our neuronal response more than a single source. Finally, recent studies have also indicated that audio/visual pairs require some correspondences to ignite facilitation. Kim et al. (2008) discovered that the auditory motion signal facilitated adults’ visual motion detection only when the auditory motion conveyed the same motion direction as did the visual motion stimuli (e.g. both visual and auditory motion signaled toward the left), but not when the auditory motion conveyed the opposite motion direction (e.g. visual motion toward the left, but auditory toward the right). Kim et al. (2012) also demonstrated that audio-visual motion congruency was critical for a task-irrelevant auditory stimulus to enhance participants’ visual motion detection – if the auditory signals conveyed the opposite direction to the visual motion display, there was no learning enhancement observed. These results suggest that bimodal learning enhancement is more than just a putative alerting effect. The visual and auditory directional information is first compared, and bimodal integration is the precursor for subsequent learning enhancement. One important note is that these accounts are not necessarily exclusive to each other. To date, there is no consensus on the mechanism underlying bimodal learning facilitation or a comprehensive study to clearly define its constraints.

To better examine the constraints and mechanisms of bimodal learning facilitation during human early development, in the current study we adopt a novel affect-congruency paradigm to investigate infants’ abstract rule acquisition. In Kim et al.’s (2012) study, the consistency of the direction of movement between audio and visual stimuli provided a good test of the origin of bimodal facilitation. Because a temporally synchronous audio-visual pair contains the same amount of visual and auditory energy as well as bimodal summation probability, any learning difference must originate from other factors. In previous infant bimodal learning studies, e.g. Frank et al. (2009) and Thiessen (2012), arbitrarily paired stimuli (i.e. geometric shape, syllables, pure tones) were chosen to constitute learning rules. These stimuli were not generated along a dimension that allows manipulation of congruency. In the current study, we created face stimuli conveying affective information in both visual and auditory dimensions, and their congruency was manipulated to tease apart possible contributing factors that govern infant bimodal learning.

Here we adopted cartoon faces with four major affective emotions – happiness, sadness, surprise, and disgust – conveyed visually with facial expressions and acoustically with matching affective voices. Affective expressions are good candidates for several reasons. First, very young infants are able to discriminate facial and vocal expressions through different affective poses such as smiling, frowning, and expressions of anger, surprise, and sadness (Barrera & Maurer, 1981; Field, Woodson, Greenberg & Cohen, 1982; LaBarbera, Izard, Vietze & Parisi, 1976; Nelson & Dolgin, 1985; Nelson, Morse & Leavitt, 1979; Young-Browne, Rosenfeld & Horowitz, 1977). Similarly, in the first 6 months of life, infants can exhibit a variety of emotions including joy, sadness, surprise, and disgust ( Bridges, 1932; Camras, 1992; Izard, Hembree & Huebner, 1987; Langsdorf, Izard, Rayias & Hembree, 1983; Stroufe, 1996; see also Lewis, 2000, for a review). Secondly, human faces and accompanying voices are ubiquitous parts of infants’ daily experiences, and emotions are universally salient and ecologically meaningful – the foundation for acquisition of social, language, and communication skills (Bahrick, 1988; Lyons-Ruth, 1977; Schiff, Benasich & Bornstein 1989; Welch & Warren, 1980). Third, infants from 4 months of age combine emotional information readily across sensory modalities (Flom & Bahrick, 2007). Thus, the relationship between affective voices and faces makes an ideal candidate for us to manipulate visual-audio congruency.

We attempted the least studied and most difficult rule (i.e. AAB) to offer additional insights on the robustness of the bimodal facilitation effect. Both behavioral and neurological studies have shown that 7-month-old infants are capable of detecting emotional incongruency between faces and voice (Soken & Pick, 1992; Walker, 1982; Walker-Andrews, 1986; Caron, Caron & MacLean, 1988; Grossmann, Striano & Friederici, 2006), so the youngest group in our study was 8-month-olds. Finally, 11-month-olds have been shown to be capable of learning AAB rules from unfamiliar visual sequences (Johnson et al., 2009), so we set the upper age limit in our testing group to 10-month-olds.

Our infants learned the target rule (AAB) when it was presented with bimodally congruent sequences (Experiment 1A and 2A). Two additional control experiments were included to show that the success in infant rule learning was not driven by visual-only (Experiment 1B) or auditory-only (Experiment 1C) affective stimuli. We further examined the origin of bimodal facilitation by mismatching the visual-audio affective pairs (Experiment 2B). Our results indicate that successful rule acquisition requires audio-visual information to be congruent. They further suggest that, during early human development,
infants actively utilize their knowledge of audio-visual congruency in learning rather than merely relying on redundancy and temporal synchrony of sensory information.

**General method**

**Participants**

We recruited 8- to 10-month-old infants in Hong Kong. All infants were full-term healthy babies with no known visual and hearing difficulties. The parents were provided with travel subsidy and a baby toy. All parents signed consent forms and were promised confidentiality.

**Apparatus**

The visual stimuli were displayed on a 19-inch ViewSonic G90fB monitor with a resolution of 1024 x 768 pixels at 85 Hz refresh rate. The average monitor luminance of the stimuli display was 66.6 cd/m². Infants’ eye and head movements were observed through a web camera and a CCTV camera, which projected the images to another monitor real-time for viewing by an experimenter blind to stimulus condition to carry out online coding. These cameras were placed right above and below the main monitor where visual stimuli were displayed. Auditory stimuli were presented over two speakers, located at each side of the main monitor. There was a screen partition between infant and experimenters to block the noises and distraction from the infant participants.

**Stimuli**

We created abstract rule sequences with affective stimuli containing emotional faces, their corresponding emotional sounds, or both.

**Visual**

Four affective cartoon facial expressions – happy, sad, surprised, and disgusted (Figure 1A) – were presented on 29 different-colored geometric shapes (default basic shapes in Microsoft PowerPoint, Version 2003) against a black background (Figure 1B). Each visual emotional stimulus started with a neutral state and gradually evolved to its full expression within 2 seconds (Figure 1A). We used cartoon affective faces to constitute the abstract rules (sample stimuli in Figure 2A, 2B). The first visual face was shown at the leftmost position on the screen for 2 s and stayed on the screen during the 2 s presentation of the second visual face in the middle of the screen. These two faces stayed on the screen during the presentation of the third face, which was added at the right-hand side of the screen for 2 s (Figure 2). The whole visual sequence stayed on the screen until the next test trial started. A simultaneous display, instead of a sequential display in which each face disappeared prior to the next face, was used to avoid working memory limitations on learning performance (Saffran et al., 2007).

**Auditory**

Each of the four emotion stimuli was presented by 10 different voices from infant, male, and female

![Figure 1](cartoon-affective-face-stimuli.png)

**Figure 1** Cartoon affective face stimuli. (A) Four basic emotions (happy, sad, surprised, disgust) were displayed to their maximum intensity for 2 seconds. (B) Twenty-nine simple shapes with different colors were used to construct affective cartoon faces conveying rules at habituation and test.
speakers lasting for 2 s. These audio files were displayed in temporal synchrony with their corresponding visual cartoon faces together forming sequences of the form AAB, ABA, or ABB.

**Design**

We adopted a within-subject design, and each infant had to complete one experiment containing two sessions, each including an habituation phase followed by a test phase. The sequence (habituation 1 – test 1 – break – habituation 2 – test 2) took about 20 minutes including a short break in the middle.

During habituation 1 and habituation 2, infants watched AAB rules. After they proceeded to the test phase (either because of meeting the habituation criteria or reaching the maximum trial limit), each infant was presented with ‘habituated AAB rule and novel ABA rule’ or ‘habituated AAB rule and novel ABB rule’. In other words, the experiment sequence is either (I) [habituation 1 (AAB) – test 1 (AAB and ABA) – habituation 2 (AAB) – test 2 (AAB and ABB)] or (II) [habituation 1 (AAB) – test 1 (AAB and ABB) – habituation 2 (AAB) – test 2 (AAB and ABA)]. The order of having novel ABA or novel ABB in test 1 was randomly assigned for each participant. Because all the shape/color and audio files at test phase were new and were never used in the habituation phase, we presented the habituated and novel rule twice each at test phase to control for the stimuli novelty effect. Infants’ average looking times recovery for the novel rules was an indicator that they could distinguish between the rules that were consistent and inconsistent with the habituated rule.

**Procedure**

For each trial, a ‘chime’ sound was used as an attention getter and the experimenter initiated each trial after the infant’s eyes and head oriented to the monitor. During the experiment, the infant sat on a parent’s lap approximately 60 cm from the monitor. Parents were instructed not to interact with their infants. An experimenter blind to the presentation of the stimuli judged whether infants’ gaze stayed on the main monitor from two hidden cameras placed right above and below the main monitor. The experimenter made real-time judgments by pressing or releasing a key to indicate whether infants’ gaze stayed on or diverted away from the monitor. The trial terminated if infants looked away from the monitor for more than 2 seconds consecutively. Infant participants were considered to have reached habituation criterion if the sum of their looking time in the last four trials declined to 50% less than of the sum of looking time of the first four trials. There was a maximum number of trials set to avoid infants becoming over-fatigued (for details please see each experiment).

**Data screening**

Because of the high cost of recruiting infant participants and huge individual variation in infants’ looking duration (Papageorgiou, Smith, Wu, Johnson, Kirkham et al., 2014), unlike the adult studies which address the outlier problem by discarding outliers or increasing sample size, it is conventional in developmental studies to set a maximum looking duration for an observational trial (e.g. Frank et al., 2009; Hamlin, Wynn & Bloom, 2007; Hamlin, Wynn & Bloom, 2010; Newman,
Herrmann, Wynn & Keil, 2008). If an infant’s looking duration exceeded this pre-set maximum duration (e.g. 30 seconds), it was recorded as the maximum duration rather than excluded as in adult studies. This method is implemented to minimize the risk that only a few individuals with extremely long looking time bias the infant group average. This approach is most effective in studies from an area with sufficient accumulation of knowledge so that a reasonable maximum duration estimate is available. In our current study, we are among the first to employ the combination of emotional face and audio in preverbal infant abstract pattern learning, and our pilot study did indicate noticeable individual variety. Due to the lack of past studies suggesting a reasonable maximum looking duration, we used two standard deviations (SD) above the group mean of infants’ looking duration at test phase (mean + 2SD) as the upper limit. This upper limit is derived separately for each experiment because our current study included bimodal and unimodal conditions, and bimodal stimulation is well-documented in adult studies to have an advantage, when compared with unimodal stimulation, in terms of neural activation (Stein & Stanford, 2008), shortened response time (Diederich & Colonius, 2004), and attentional allocation (Driver & Spence, 1998). It is reasonable to speculate that infants’ looking duration in bimodal and unimodal conditions may also vary due to the nature of stimuli across experiments. Thus, we adopt experiment-based upper limits to lower the impact from outliers without having to exclude infants who were attentive throughout the whole experiment.

**Experiment 1A: Congruent bimodal affective stimuli**

We exposed fifteen 8- to 10-month-olds (5 male, mean age = 285 days, $SD = 16.48$ days) to simultaneous facial expressions defined by geometric shapes along with the corresponding affective sounds (Figure 1, Figure 2A). Three additional infants (at mean age of 284.67 days) were excluded due to fussiness (2) and lack of attention (1).

In Experiment 1, all infants watched the AAB rule repetitively until their looking time in the last four trials declined to 50% less than the sum of looking time of the first four trials. To avoid fatigue, we set 12 trials as the maximum number of habituation trials as in previous studies (e.g. Frank et al., 2009; Johnson et al., 2009), and any infants who did not habituate within 12 trials proceeded directly to the test phase as in previous studies.

**Results**

The average number of trials to reach habituation was 9.8. The looking time to novel (ABA or ABB rule) and habituated rules (AAB rule) at test phase is depicted in Figure 3. The mean looking time per trial was 5.9859 seconds; $SD$ 3.449 seconds. Infants’ looking time in three trials exceeded the upper limit (12.88 seconds) and therefore was replaced with 12.88 seconds.

Each infant participated in two sessions of the experiment (novel rule type either ABA or ABB), and in each session there were two types of rule (familiar or novel rule) with two trials for each rule type. There were four conditions altogether (2 rule types: familiar or novel × 2 novel rule types: ABA or ABB), and two trials for each. We computed an average of the looking time of the two test trials for each condition. As all infants (after exclusion) participated in all four conditions, we entered the four numbers (average looking time for the four conditions) into a $2 \times 2$ repeated measures analysis of variance (ANOVA) with the following two factors: Novel rule (ABA vs. ABB) and trial type (habituated vs. novel rule). Infants looked significantly longer at novel rules than the habituated AAB rule ($F(1, 14) = 36.756$, $p < .001$), suggesting that they discriminated the novel patterns from the habituated rule. Altogether, 13 of 15 infants looked longer at the ABA rule, which was significantly above chance ($\chi^2(1) = 8.067$, $p = .005$).

There were no significant differences in Novel rule and interaction effects. For the eight (in the ABA rule) and nine (in the ABB rule) infants who reached the maximum habituation trials, further analysis reported that their looking time difference between novel and learned rule at test stage was not different from those who were successfully habituated [ABA rule, $t(13) = 1.014$, $p = .329$; ABB rule: $t(13) = 0.857$, $p = .407$].

In summary, infants demonstrated a preference for the novel rules as opposed to the habituated rule, suggesting that they discriminated between the habituated rule and novel rules.

**Experiment 1B: Visual affective stimuli only**

As a control experiment, we tested whether the 8- to 10-month-olds were able to learn the same rules from visual affective stimuli alone. We tested fourteen 8- to 10-month-old infants (10 male, mean age = 283.57 days, $SD = 16.28$ days) with the same method as in Experiment 1A, except that we removed the affective sounds. Three additional infants at a mean age of 283.57 days were excluded due to fussiness (1), sleepiness (1) and lack of attention (1).
Results

The average number of trials to habituation was 10. The results are summarized in Figure 3. The mean looking time per trial was 5.645 seconds; SD was 3.815 seconds. This resulted in the upper limit (mean + 2SD) being set at 13.27 seconds, and four outliers exceeding this value were replaced with 13.27 seconds. We entered all data of test trial looking times into a 2 (test trial type: habituated rule vs. novel rule) × 2 (novel rule type: ABA vs. ABB) repeated measures ANOVA as in Experiment 1. We did not find any main effect in either factor or interaction (Fs < 1, ps > .5). Seven of 14 infants showed novelty preferences in both ABA and ABB condition, which was not significantly different from chance ($\chi^2 (1) = 0.00, p = 1$).

There were six (in the ABA rule test) and seven (in the ABB rule test) infants who did not reach habituation criteria and they proceeded to the test phase after reaching the maximum number of trials. Their test performance was no different from those who met the habituation criteria [ABA rule, $t(12) = 0.048, p = .962$; ABB rule, $t(12) = -0.37, p = .718$].

The results in Experiment 1B suggested that the infants were not able to acquire the AAB rule from visual affective stimuli alone, rejecting the hypothesis that the rule learning of a difficult rule (AAB) in Experiment 1A was only from affective visual stimuli.

Experiment 1C: Auditory affective stimuli only

In Experiment 1C, we examined fifteen 8- to 10-month-olds (9 male, mean age = 268.93 days, SD = 17.89 days); five additional infants at a mean age of 259 days were excluded due to sleepiness (2), lack of attention (2), or impatience (1). We reused 32 out of 40 auditory stimuli from Experiment 1 (eight were excluded because in a pilot test our adult subjects found it difficult to distinguish affect with auditory stimuli alone). An eye rolling in the center of the grey screen served as the fixation point, and infants’ looking time toward any part of the monitor was recorded as in the previous two experiments. All other procedures and methods were identical to those in Experiment 1A.

Results and discussion

The average number of trials to habituation was 9, and the test stage looking time is summarized in Figure 3.

The mean looking time per trial was 4.323 seconds; SD was 2.685 seconds. This resulted in the upper limit (mean + 2SD) being set at 9.69 seconds, and three outliers exceeding this value were replaced with 9.69 seconds. We entered all the data of the test trial looking times into a 2 × 2 repeated measures ANOVA with the following two factors: test trial type: (habituated rule vs. novel rule) × 2 (novel rule type: ABA vs. ABB).
novel rule) and novel rule type: (ABA vs. ABB). We did not find any main effects in test trial rule, trial type, or the interaction (Fs < 1, ps > .5). In the ABA and ABB test conditions, 8 out of 15 and 7 out of 15 infants, respectively, showed novelty preferences, which was not significantly different from from chance [$\chi^2(1) = 0.067$, $p = .796$ for ABA test; $\chi^2(1) = 0.067$, $p = .796$ for ABB test]. For those who were not successfully habituated (five in the ABA condition and two in the ABB condition), further analyses confirmed that their performance at test did not differ from those who were habituated ($t < 1$, $p > .5$). Thus, all the data were included in the analyses.

The results of Experiment 1C also suggested that the infants were unable to learn the AAB rule from auditory affective sound alone, rejecting the conjecture that the rule learning in Experiment 1A was based on the infants’ ability to encode rules from auditory stimuli alone.

Cross-experimental analysis (Experiments 1A, 1B, and 1C)

Because the number of infants meeting the habituation criteria varied in the three experiments, we did a subsequent cross-experimental analysis to test the hypothesis that bimodal facilitation is a result of the higher number of habituated infants in Experiment 1. We conducted a 2 (habituation status: habituated or not habituated) x 3 (stimuli types: bimodal/visual/auditory) ANOVA on the looking time difference between habituated rule and the novel rules at test phase. If bimodal facilitation (in Experiment 1A) is a result of more effective habituation, we should expect a significant interaction effect between these two factors. We found no evidence suggesting that habituation status differed across experiments, and our results suggested that stimuli type was the only significant factor [$F(2, 82) = 8.421$, $p < .001$], with no main effect on habituation status nor any interaction effects.

Discussion of Experiment 1

We discovered that non-linguistic bimodal affective sequences facilitated infant rule acquisition. As mentioned in the introduction, our results in Experiment 1A might originate from several potential factors. Firstly, redundant sensory inputs might enhance learning because they double the statistical occurrence (i.e. bimodal is more than unimodal). Secondly, bimodal sensory stimulation might arouse infants’ alertness or alter selective attention (i.e. Intersensory Redundancy Hypothesis) and thereby promote learning (see Bahrick & Lickliter, 2012). Thirdly, the relational congruency of bimodal stimuli might provide a stronger and effective sensory unit for rule extraction and subsequent learning.

Experiments 1A to 1C serve as a foundation block for us to further disentangle the above-mentioned possibilities. In Experiment 2, we mismatched the emotion relationship between audio-visual pairs so that audio and visual stimuli were relationally incongruent.

If the learning effect persists, it will demonstrate that redundant and synchronous sensory information is sufficient to promote bimodal facilitation. Otherwise, infant learning involves bimodal integration beyond temporal synchrony matching.

Experiment 2A: Bimodal congruent affective stimuli

Because our infant lab moved to a new space we first replicated the previous Experiment 1A in the new environment. In Experiment 1, we had a high proportion of infants (over 50%) who reached the maximum 12 trials without meeting the habituation criteria, so we took a more conservative ‘habitation paradigm’ in Experiment 2. We only included infants who reached habituation in our data analysis. The experiment was stopped if an infant failed to reach the habituation criterion within 20 trials to avoid them becoming over-fatigued at the subsequent test phase. The other stimuli and testing procedures were identical to Experiment 1A.

Participants

Fifteen infants were tested in this experiment (10 male, mean age = 285.1 days, $SD = 14.89$ days). Four additional infants were excluded due to fussiness (2) or unsuccessful habituation (2) and their average age was 287.5 days.

Results

The average number of trials to habituation for this experiment was 9.6. The mean looking time of test trials was 9.710 seconds; $SD$ was 6.925 seconds. This meant that the upper limit (mean + 2SD) was 23.56 seconds, and three outliers exceeding this value were replaced with 23.56 seconds.

The looking time results in the test phase are depicted in Figure 3, and they were entered to a 2 x 2 repeated measures ANOVA with the following two factors: novel rule type (ABA vs. ABB) and test trial type (habituated
rule vs. novel rule). Only the main effect of test trial type was significant \((F(1, 14) = 21.126, p < .001)\), whereas the main effect of novel rule type and the interaction were not \((ps > .5)\). A chi-square goodness of fit test showed that the number of infants who looked longer at novel rules was significantly more than the number expected by chance \((13\text{ out of 15 demonstrated novelty preference in the ABA test condition: } X^2(1) = 8.07, p = .0045; 14\text{ out of 15 infants showed novelty preference in the ABB test condition } X^2(1) = 11.27, p = .0008)\).

In this experiment, we replicated the results in Experiment 1A: infants were able to discriminate between habituated AAB rule and novel ABA and ABB rules.

**Experiment 2B: Incongruent bimodal affective stimuli**

In Experiment 2B, we tested whether mismatched emotional pairings would limit infant rule acquisition. All the procedures were identical to Experiment 2A, except that the pairing of audio-visual stimuli was mismatched (for details see below).

**Participants**

Seventeen infants were tested in this experiment (7 male, mean age = 283.90 days, \(SD = 22.56\) days). Nine additional infants were excluded due to fussiness (3), unsuccessful habituation (3), or incompletion of the task (3), and their average age was 284.8 days.

**Stimuli**

The visual and auditory affective stimuli were identical to those used in Experiment 2A except that their emotional contents were mismatched (incongruent) during presentation. Figure 2B depicts an example in which a happy face was paired with a surprised voice.

**Results**

The average number of trials to habituation was 9.7. The mean looking time per trial at test phase was 12.970 seconds; \(SD = 11.624\) seconds. This meant that the upper limit (mean + 2SD) was 36.21 seconds, and three outliers exceeding this value were replaced with 36.21 seconds. The looking times to novel (AAB or ABB rule) and habituated rules (AAB rule) at test stage were entered into a 2 (test trial type: habituated rule vs. novel rule) \(\times\) 2 (novel rule type: ABA vs. ABB) repeated measures ANOVA. Figure 3 shows the looking time results.

Neither the main effects nor the interaction were significant \((Fs < 1, ps > .5)\). A chi-square goodness of fit test showed that the number of observers who looked longer at novel rules was not significantly more than the number expected by chance \((10\text{ out of 17 demonstrated novelty preference in ABA test condition: } X^2(1) = 0.529, p = .47; 12\text{ out of 17 demonstrated novelty preference in ABB test condition: } X^2(1) = 2.88, p = .09)\). Thus, with incongruent affective bimodal stimuli, infants are less able to differentiate a habituated rule from a new rule.

**Discussion of Experiment 2**

The relational congruency between audio and visual emotion is therefore critical for infants to extract the invariant rules embedded under the sequence display. Because Experiments 2A and 2B were both bimodal, the rule acquisition advantage in Experiment 2A (over Experiment 2B) did not originate from pure sensory enhancement or bimodal statistical summation. Rather, the results imply that infants process each bimodal unit individually before connecting the temporal sequence into a rule pattern. In Experiment 2B, although the corresponding visual and auditory emotions were mismatched, each modality itself constituted the identical AAB rule. If infants had extracted the AAB pattern separately from the auditory stream (voice) or visual stream (face) alone first, then bimodal integration occurred, we would have expected infants to exhibit at least the same amount of rule acquisition as in those in Experiment 2A. Thus, our finding demonstrated that infants integrate bimodal information first before applying these bimodal units in rule learning.

**General discussion**

The current findings enrich our understanding of bimodal facilitation learning effects by demonstrating that 8- to 10-month-olds are only able to acquire the more difficult rule (AAB) from sequences with congruent bimodal affective stimuli presentation. When the same rule was presented with unimodal affective stimuli alone or under incongruent bimodal affective stimuli (i.e. laughing face paired with crying sound), infants failed to acquire the same underlying patterns. This leads us to conclude that bimodal facilitation in infant rule learning is not the sole product of sensory redundant inputs across multiple modalities.

Rule learning is a complicated cognitive ability involving multi-stage processing. Enhancement at any stage could facilitate the rule acquisition. In which stage is bimodal facilitation most likely to occur?
A number of studies have suggested that adults’ perceptual experience can be enhanced when information from different modalities is congruent. Chen, Yeh and Spence (2011) found that visual images presented congruently with audio contents received higher processing priority. When the observer’s two eyes were presented with two different images (e.g. a bird to the right eye, and a car to the left eye), the two images would compete to be manifested as the dominant percept (i.e. binocular rivalry). If bird sounds were displayed, the corresponding congruent bird image would be more likely to be reported by observers as the perceived dominant percept than when the background audio was an irrelevant or incongruent sound, such as a restaurant or car sounds. Hsiao, Chen, Spence and Yeh (2012) presented observers with an ambiguous image that could be perceived either as an old or young female (i.e. bistable figure). They found that when the voice of an old lady was presented (congruent), observers were more likely to perceive the ambiguous image as an old lady than when a young lady’s voice was presented (incongruent), or when a beep sound was presented (irrelevant). Interestingly, when participants were asked to attend to one form of the bistable image (e.g. attend to the old lady’s image), hearing incongruent audio (e.g. a young lady’s voice) significantly lowered the chance of perceiving the targeted form of the image (i.e. old lady’s image) than hearing an irrelevant beach waves sound or no sound. The influence of congruency reflected participants’ knowledge of audio-visual associations in daily experience and also participants’ semantic interpretation (e.g. woman from young/old age group, bird/car sounds).

Infants’ perceptual encoding also benefits from audio-visual congruent relationships. Srinivasan and Carey (2010) used synchronous visual cartoon caterpillars with varying length (spatial dimension) and auditory tones with different duration (temporal dimension) to investigate whether 9-month-old infants can associate representations from visual/spatial with auditory/temporal dimensions. They found that infants were able to associate positively related pairs (i.e. long caterpillar was paired with long duration tone) but not negatively related pairs (i.e. short caterpillar was paired with long duration tone). A recent ERP study conducted by Hyde, Porter, Flom and Stone (2013) replicated Srinivasan and Carey’s study (2010) with 5- to 6-month-olds found that relational congruency may promote early processing of sensory information and heighten attention during familiarization, leading to better information encoding. Taken together, the above studies with adults and infants all illustrate that relational congruency under bimodal stimulation can facilitate subsequent learning by strengthening the encoding stage.

The bimodal learning facilitation could originate from better attentional allocation as well. The Intersensory Redundancy Hypothesis (IRH) is one prominent theory that suggests that bimodal conditions affect infant perception through attentional selection (see Bahrick & Lickliter, 2012, for a review). According to this theory, infants prioritize information presented concurrently and synchronously across multiple modalities (intersensory redundancy) over unimodally presented information. Its testable predictions have been well supported by behavioral studies (Bahrick & Lickliter, 2000, 2002; Bahrick et al., 2004; Lickliter and Bahrick, 2004; Lewkowicz, 2000; Bahrick, 2010). However, IRH focuses on how infants’ selective attention is directed by unimodal or bimodal stimulation, and little emphasis is put on the constraints between the cross-modal relations. Our affective congruent and incongruent conditions contained bimodal inputs, and the audio sequence and visual sequence alone in both conditions conveyed the same principle. Under the framework of IRH, they are not expected to differ. However, whether congruency between two sensory modalities alters viewers’ attentional state is still an open question that requires further investigation.

An important contribution of our current paper is that we demonstrated the role of top-down factors in early infant multi-sensory integration and learning. Both bottom-up (or structural) and top-down (or cognitive) factors interact with multi-sensory integration (Chen & Spence, 2010). The former includes stimuli-driven elements such as temporal synchrony and location proximity, likely driving the bimodal learning facilitation observed in Frank et al. (2009) and Thiessen (2012). The latter involves our previous knowledge of the incoming sensory stimuli such as the knowledge of matching affective sounds with affective facial expression in our study. Our tested infants (8–10 months old) are older than those in Frank et al. (2009) and Thiessen (2012), and a natural follow-up question is to ask whether infant maturity governs how bimodal stimulation benefits their learning. This is similar to the central view of multi-sensory perceptual narrowing (Lewkowicz & Ghazanfar, 2009; Lewkowicz, Sowinski & Place, 2008; Lewkowicz, 2010), which suggests that when infants grow, their experiences direct them to use higher-level perceptual cues rather than lower-level perceptual cues to learn a multi-modal event. This is another direction for future research.

In summary, our findings add to the converging literature that successful bimodal learning requires audio-visual relational congruency. It reveals that young infants do not merely rely on perceptually redundant information under bimodal stimulation, but they also actively employ their current knowledge to interpret audio-visual information during the learning processes.
Acknowledgements

This work was supported by the General Research Fund from Research Grants Council of Hong Kong, China, and Hong Kong University Seed Funding Programme for Basic Research to Dr Chia-huei Tseng. We thank Diane Yat, Ping-hui Chiu, Geroldene Tsui, Stephanie Chen and Fiona Hui for their assistance in data collection.

References


