AUDIOVISUAL SPEECH INTEGRATION IN INFANCY: EVIDENCE FROM THE MCGURK EFFECT IN A PERCEPTION-BASED BEHAVIOURAL TASK

RACHEL XINYANG LIU, B.A., M.A.

A Thesis Submitted to the School of Graduate Studies in Partial Fulfilment of the Requirements for the Degree Master of Science

McMaster University MASTER OF SCIENCE (2025) Hamilton, Ontario (Psychology, Neuroscience and Behaviour)

TITLE:	Audiovisual Speech Integration in Infancy: Evidence from the McGurk Effect in a Perception-Based Behavioural Task
AUTHOR:	Rachel Xinyang Liu
SUPERVISOR:	Dr. Gabriel Naiqi Xiao
PAGES:	CX, 110

Abstract

Multisensory speech perception plays a critical role in language acquisition and later socio-cognitive and socio-emotional development. Recent evidence suggests that infants are able to integrate auditory and visual speech information early in life, and that this integration capacity is modulated by early visual experience with the familiar-race faces. Using the McGurk effect as an index of visual speech influence on auditory perception, we tested infants aged 6 to 12 months to examine (1) whether they could integrate audiovisual speech cues and whether this capacity strengthens with age, and (2) whether this integration is modulated by face-race familiarity. Infants participated in a perception-based behavioural task modeled on the Stimulus-Alternation Preference Procedure (SAPP; Best & Jones, 1998), and their looking times in response to McGurk and non-McGurk syllable pairs enacted by own-race and other-race faces were recorded using eye-tracking. When viewing own-race faces, infants demonstrated a robust and stable audiovisual integration capacity during the tested age range. However, other-race faces significantly disrupted this capacity from 6 months onward, indicating an early-emerging other-race effect (ORE). These findings support the view that speech is presented bimodally and processed in an integrative manner from early in life, and that perceptual tuning is a cross-modality, pan-sensory phenomenon. They further contribute to the argument that multisensory perceptual development involves a regressive reorganization that calibrates infants' perceptual systems to the most ecologically relevant information.

Acknowledgements

I would love to express my heartfelt gratitude to those who have shaped who I am as a researcher today. First and foremost, thank you to my supervisor, Dr. Gabriel N. Xiao, who introduced me to the field of perception and cognition in infancy, provided a transformative environment that inspired my intellectual curiosity and discovery, and guided me through the past two years of my academic journey. I am truly grateful for your unwavering support and empowering guidance. The discussions we have had about literature, experiments, and research skills will forever stay with me and continue to propel me along my future research path.

I would also like to thank my committee members, Dr. Janet F. Werker and Dr. Laurel J. Trainor. I first came to know you while reading literature that sparked my interest and lit up my passion for the scientific study of infancy—and I am beyond fortunate to now have you both as mentors in research and in life. Your insightful feedback, kind encouragement, and thoughtful, friendly challenges helped me navigate obstacles and always inspired me to grow "a head taller." You both exemplify the brilliance of women scientists, and I aspire to become one like you.

To my lab mates—Vivian Fang, Carie Guan, and Anagha Vinod—thank you all for always being open to trying out new ideas together and bringing laugher to the lab. A special thank-you to all the participating families, some of whom drove hours to the Baby Lab. Your support is vital to the research taking place at McMaster Infant Studies Group, including my own, and it contributes so meaningfully to our understanding of infancy. I am also deeply grateful to my wonderful RAs—Anushka Goyal, Alsana Amrenova, Maria Dondas, and Thuy Tran, among others—for your dedication to recruitment and your help with the study sessions.

Last but not the least, thank you to my family for your irreplaceable and unconditional love, support, and understanding—they are everything I could ever ask for in my entire life.

Table of Contents

List of Figures

List of Tables

Introduction

The world around us abounds with sensory signals from multiple modalities. The continuous influx of these signals necessitates an adept capacity to perceive and integrate them in order to construct a coherent perceptual experience (e.g., Lee & Wallace, 2019; Lewkowicz & Ghazanfar, 2009). As one of the most crucial sources of information in everyday face-to-face communication, talking faces emit a wealth of multisensory inputs that jointly shape our perception of speech. Specifically, when watching and listening to someone talk, we process (1) modality-specific attributes unique to each sensory modality, including visual cues such as the interlocutor's racial identities and facial feature movements, as well as auditory cues such as speech segments and prosodic features; (2) spatiotemporally congruent information comprising signals that occur in synchrony and originate from the same location, such as seeing a moving mouth while hearing the corresponding speech; and (3) invariant amodal attributes—such as temporal synchrony (Bahrick & Hollich, 2008)—which remain consistent across modalities and index the coordination between visible and audible articulatory actions (e.g., Chandrasekaran et al., 2009). Therefore, speech perception is an inherently multimodal event composed primarily of audiovisual information (Rosenblum, 2008), and our ability to effectively integrate these multisensory inputs is essential for deriving a systematic and reliable perceptual entity from the redundant sensory information (Bahrick et al., 2004; Lewkowicz & Kraebel, 2004).

Although adults are well-equipped with this multisensory speech integration ability (e.g., McGurk & MacDonald, 1976), one of the most daunting tasks facing infants is to discover the multisensory coherence of the speech events that make up the perceptual ecology of their everyday environment. From the moment of birth, talking faces and their accompanying auditory and visual cues pervade infants' multimodal experiences, especially through interactions with their caregivers (Gijbels et al., 2025; Lewkowicz, 2010). While essential for early language

development and gradual socialization into their sociolinguistic communities (e.g., Altvater-Mackensen & Grossmann, 2015), as well as later socio-cognitive and socio-emotional development (Bahrick & Lickliter, 2012; Guiraud et al., 2012; Wallace et al., 2020), audiovisual speech integration poses considerable challenges to infants due to an immature perceptual system.

Nevertheless, empirical evidence suggests that infants' multisensory perceptual abilities undergo rapid development over the first year of life—a trajectory explained by three competing theoretical accounts. The first posits that intersensory perception is minimal at birth and gradually *emerges* through experience as infants learn to bridge unimodal sensory information together (Birch & Lefford, 1967; Piaget & Cook, 1952). The second contends that infants are born with rudimentary multisensory abilities that are initially unified but become increasingly differentiated through exposure to more refined stimulation (Gibson, 1984). The third account proposes that infants are equipped with a broadly tuned multisensory capacity from birth, yet this capacity undergoes a reorganization during the first year of life to align with the infants' specific environment—a process commonly referred to as perceptual tuning or narrowing. This reorganization hones a more focused perceptual system that readily processes the most relevant information in their surroundings (Lewkowicz & Ghazanfar, 2009). While the first two accounts emphasize a progressive nature of multisensory perceptual development, the third highlights its regressive nature—arguably preparing infants with a perceptual system that is optimized for efficiently learning about their immediate environment (e.g., Hay et al., 2015).

Based on the belief that multisensory speech perception is present at birth (see *Bimodal nature of speech representation in early infancy* in the following review), of interest in the present study is whether the second and third account applies to the development of multisensory

speech perception during infancy. In particular, we sought to examine infants' ability to *integrate* auditory and visual speech information to form a unified perceptual experience, and whether early experience—largely shaped by interactions with own-race individuals—influences this capacity. Although this issue has been studied previously, two persistent problems have limited interpretability in this line of research: a lack of terminological precision, which conflates infants' detection of audiovisual coherence with temporal synchrony, as well as methodologically varied experimental designs and stimuli (see Lozano et al., 2024; Shaw & Bortfeld, 2015, for a review). To address these issues and improve conceptual and methodological clarity, the present study tested speech integration at the phonetic level using the well-documented McGurk effect (McGurk & MacDonald, 1976), with auditory and visual speech information always temporally aligned. Additionally, we designed a perception-based behavioral task modeled on the Stimulus-Alternation Preference Procedure (SAPP; Best & Jones, 1998) and paired it with a gaze-contingent stimulus presentation to measure infants' real-time perception in a sensitive manner.

Before introducing the experimental details of our study, we provide a brief overview of the broader topic of audiovisual speech perception in infancy. This includes (1) the early and robust bimodal representation of speech, (2) evidence of early audiovisual speech integration, and (3) evidence that such bimodal integration capacity might not be robust in infancy—from the perspectives of a continuous increase of visual speech influence beyond infancy and the disruptions on integration induced by other-race faces, an unfamiliar social visual category.

Bimodal nature of speech representation in early infancy

One crucial prerequisite for integrating auditory and visual speech information is the recognition that speech is intrinsically bimodal. Notably, this understanding emerges prior to the

onset of spoken language, as evidenced by prelinguistic infants' remarkable sensitivity to the linkage between auditory and visual speech information emanating from a speaker's mouth.

Kuhl and Meltzoff (1982) were the first to uncover infants' early ability to detect alignments between articulatory mouth movements and vowel sounds. Using the preferential looking technique, they presented 4.5-month-olds side-by-side images of a female articulating /i/ (widespread lips) and /a/ (open jaw) while simultaneously playing an auditory vowel that corresponded to only one of the two visual articulations. Infants looked significantly longer to the audio-visually matched face than would be expected by chance, suggesting a sensitivity to the bimodal representation of speech that allows them to associate phonetic information conveyed by lip movements (i.e., visual articulatory cues) with that conveyed by the voice (i.e., acoustic speech cues). This audiovisual matching ability at 4.5 months of age was further replicated with a different group of same-aged infants (Kuhl & Meltzoff, 1984) and with a new vowel contrast, /i/ and /u/ (rounded mouth) (Kuhl & Meltzoff, 1988). The robustness of such intermodal representation of speech was substantiated when same-aged infants successfully connected auditory vowels with their corresponding articulatory movements on *male* faces (Patterson & Werker, 1999), a less familiar and less preferred gender category to infants compared with females, as confirmed in both experimental (e.g., Quinn et al., 2002) and naturalistic contexts (Sugden et al., 2013).

A series of follow-up studies explored the developmental origins of this bimodal representation of speech—also using the phonetic matching effect as a proxy—and found it to be stable enough to emerge with minimal postnatal linguistic experience. Using an operant-choice sucking procedure—which enables infants to modulate their sucking to control which stimulus they wish to inspect—and the vowel pair /i/ and /u/, Walton and Bower (1993) found that 4-

month-olds preferred viewing faces with lip movements that corresponded to the concurrently presented auditory vowel. They further argued that this ability was not driven by linguistic familiarity, as 6- to 8-month-old English-learning infants sucked more to view an unfamiliar but congruent vowel-face pairing (auditory French /y/–visual English /u/) than a mismatched one (auditory English /i/–visual /u/). Since the vowel phoneme /y/ is non-native, the researchers concluded that infants' preference for matched articulatory and acoustic information was not attributable to learning or language familiarity, but rather, reflects a biologically predisposed sensitivity. To further test this possibility, Patterson and Werker (2003) extended the investigation to even younger infants with a preferential looking design. They observed that 2-month-olds looked significantly longer at both female and male faces articulating vowels that aligned with a concurrently heard vowel (/i/ or /a/), showing a clear-cut audiovisual mapping comparable to that previously observed in 4.5-month-olds (Patterson & Werker, 1999). Except for vowels, infants aged 2 to 4 months can also match audible and visual consonants (MacKain et al., 1983).

Converging with these behavioural findings, neurophysiological evidence yields compelling support for a hardwired audiovisual representation of speech. Bristow et al. (2009) employed high-density event-related potential (ERP) recordings in a mismatch paradigm, in which a deviant stimulus is introduced following repetitions of a standard stimulus, with 10-week-old infants. After being familiarized with a silent video of a speaker articulating either /a/ or /i/, infants were presented with an auditory vowel that either matched or mismatched the previously seen visual articulation. Incongruent auditory stimuli elicited a mismatch response (MMR)—a neural signal typically associated with the detection of violations in predicted sensory inputs. Notably, this MMR triggered by cross-modal mismatch closely resembled that elicited by unimodal auditory mismatch, both in timing and scalp distribution, and was localized

to left-lateralized fronto-temporal regions associated with phonetic processing. These findings suggest that, as early as two months of age, visual speech cues are encoded in a phonetic format that can be directly compared with auditory inputs—supporting the view that infants possess a structured, bimodal representation of speech from a remarkably young stage in development.

However, since the participants for all the abovementioned studies had already received some—albeit arguably not extensive—postnatal linguistic exposure by the time of testing, it remains inconclusive whether phonetic speech perception is innately bimodal. To address this, Albridge et al. (1999) measured English-immersed newborns aged 4 to 33 hours with the operant choice-preference sucking procedure. The stimuli included matched and mismatched audiovisual presentations combining the acoustic and articulatory forms of four vowel phonemes: three native (/i/ with widespread lips, /a/ with an open jaw, and /u/ with rounded lips) and one non-native (French /y/), which, similar to the English /u/, is also pronounced with rounded lips. The results showed that newborns preferred matched face-voice pairings (e.g., rounded lips dubbed with auditory /u/) over mismatched ones (e.g., rounded lips dubbed with auditory /a/), a preference that extended even to stimulus pairings involving the non-native phoneme /y/. Given that the French /y/ does not exist in English, to which the newborns could have been exposed to solely in auditory forms in prenatal life (Kisilevsky et al., 2009; Partanen et al., 2013), the authors suggested that the observed preference indicates a built-in structural sensitivity to visual speech (e.g., 'innately guided' in Jusczyk & Bertoncini, 1988).

In addition to studies that directly examined infants' cross-modal phonetic equivalence, several findings on infants' articulatory imitations of a speaker's vocalizations provide incidental evidence of early sensitivity to visual speech. For instance, 3-month-olds in Legerstee (1990) produced more self-initiated vocal imitations of /i/ and /a/ when these vowels were presented in

an audiovisually congruent than incongruent manner. Similar mimicry patterns in response to matched phonetic information across modalities were also observed in 4.5-month-olds (Kuhl & Meltzoff, 1988, 1996), regardless of the speakers' gender familiarity (Patterson & Werker, 1999).

In sum, infants' early structural sensitivity to audiovisual speech correspondences—demonstrated by their effortless matching of speech information presented in face-voice pairs and their spontaneous vocal imitation of congruent pairings—suggests that speech is represented bimodally from birth, requiring only minimal exposure to talking faces. This intermodal sensitivity appears particularly robust at the *phonetic* level, emerging earlier than other forms of face-voice matchings involving social attributes such as age (by 7 months; Bahrick et al., 1998), gender (by 6–8 months; Walker-Andrews et al., 1991; Patterson & Werker, 2002), and emotion (by 5 months; Walker, 1992), implying that phonetic-level speech cues may engage lower-level perceptual mechanisms relative to the processing of other socially salient facial features.

Evidence of early audiovisual speech integration

Having established that infants possess innate knowledge that speech is bimodally represented, which enables them an early sensitivity to visual speech to reliably *match* phonetic information across auditory and visual modalities, an important next step is to consider how this foundational ability sets the stage for more complex speech processing mechanisms. Building on this, amounting evidence suggests that infants do not simply detect equivalence between auditory and visual speech events; rather, they also *integrate* them into a cohesive perceptual entity, treating them as complementary sources of communicative information that can be encoded within a shared representational space.

The McGurk effect as an index

An ideal tool for examining this audiovisual speech integration capacity is *the McGurk effect*, a well-known perceptual illusion that exemplifies the influence of visual speech cues on

auditory perception (McGurk & MacDonald, 1976; Rosenblum et al., 1997). It serves as a valuable proxy for investigating audiovisual speech integration capacity that enables more coherent and meaningful perceptual experiences of speech events. An example of this illusion is that when hearing the syllable /ba/ dubbed onto lip movements producing /ga/ (i.e., auditory /ba/-visual /ga/), people would hear an *intermediate*[1] phoneme /da/; similarly, when hearing an auditory /pa/ while seeing a mouth that synchronously articulates /ka/ (i.e., auditory /pa/-visual /ka/), people often perceive an *intermediate* /ta/. These fused responses indicate that the simultaneous presentations of the incongruent phonetic and visual cues are processed interactively, generating a unified percept distinct from the physical stimuli in either of the two sensory modalities (Beauchamp et al., 2004; Nath & Beauchamp, 2012).

However, not all mismatched audiovisual combinations elicit such illusory fusions. For instance, reversals of the auditory and visual speech syllables in the McGurk audiovisual pairings (e.g., auditory /ga/-visual /ba/ and auditory /ka/-visual /pa/) typically result in the perception of the original auditory—or even visual—syllable, or yield *combination responses* such as /bga/ or /pka/, which preserve relatively unaltered stimulus properties from both modalities (Kushnerenko et al., 2008; Kushnerenko et al., 2013; McGurk & MacDonald, 1976). Interestingly, the perceptual outcomes of these "non-McGurk" pairings tend to remain aligned with the physical auditory syllables in younger children but become increasingly influenced by the visual syllables in adults (McGurk & MacDonald, 1976), as will be discussed in greater details in the *A protracted developmental trajectory of AV-speech integration* section.

[1] Intermediate in the sense of place of articulation, a concept that will be detailed shortly. Briefly, the physical auditory stimulus /ba/ starts with a bilabial consonant, articulated with both lips at the very front of the oral cavity. The physical visual stimulus /ga/ begins with a velar consonant, pronounced by contacting the tongue body with the velum at the very back of the oral cavity. The fused percept /da/ starts with an alveolar consonant, articulated at a mid-oral cavity location where the tongue tip contacts the alveolar ridge. As such, /da/ occupies an intermediate articulatory position between /ba/ and /ga/. The same logic applies to the auditory /pa/-visual /ka/ pairing.

The underlying mechanisms for fused and combination responses are phonological. Specifically, the perceptual outcomes are shaped by the places and manners of articulation of the stimulus auditory and visual syllables, especially the former. *Place of articulation* refers to which articulators (i.e., speech organs) are involved and where they are positioned to produce a particular consonant (e.g., bilabial, alveolar, velar), thereby influencing the visual appearance of a speech sound. *Manner of articulation*, on the other hand, describes how airflow is manipulated to produce a consonant (e.g., stop, fricative), thus determining its acoustic features (Celce-Murcia et al., 2010). Together, these two phonetic dimensions decide whether the incongruent auditory and visual syllables can be integrated into a unified third percept.

In McGurk pairings that typically yield integrated percepts—such as auditory /ba/-visual /ga/ (fused as /da/) and auditory /pa/-visual /ka/ (fused as /ta/)—the visual syllables /ga/ and /ka/ are *velar* stops produced by contact between the tongue body and the velum, the soft palate at the very back of the oral cavity. In contrast, the fused percepts /da/ and /ta/ are *alveolar* stops produced near the middle of the oral cavity, where the tongue tip contacts the alveolar ridge. Despite differing in the precise articulatory locations, velars and alveolars share a similar visible articulatory gesture—a slightly open mouth—since both are produced inside the oral cavity. This visual similarity may lead to perceptual confusion, contributing to the emergence of a fused percept (Lindborg et al., 2021; Tiippana et al., 2023; Van Wassenhove et al., 2005). Furthermore, the stimulus auditory syllables /ba/ and /pa/ are *bilabial* stops, with /ba/ being voiced and /pa/ voiceless. Their corresponding fused percepts—/da/ and /ta/—match these features in terms of voicing and manner. More importantly, the relatively close places of articulation between the auditory syllables (i.e., lips) and their fused auditory percepts (i.e., the alveolar ridge) create similar oral resonance, reinforcing their acoustic similarity. Taken together, the conflicting

auditory and visual cues guide the perceptual system toward a single, integrated syllable that visually resembles the visual input and is acoustically similar to the auditory input.

In contrast, the absence of illusory fusion in the non-McGurk pairings can be attributed to greater phonological incompatibility between the auditory and visual cues. In pairings such as auditory /ga/-visual /ba/ and auditory /ka/-visual /pa/, the visual syllables /ba/ and /pa/ are bilabial consonants produced by bringing both lips together. These gestures are highly visually salient and align exclusively with bilabial auditory consonants (i.e., /b/, /p/, /m/) because lips are the outermost articulator. However, the auditory syllables /ga/ and /ka/ are velar consonants articulated at the back of the oral cavity without any lip involvement. Another key consideration is the lack of plausible substitutes for the auditory velars /ga/ and /ka/, as the English phonetic inventory does not include phonemes produced at the hard palate—the articulatory region adjacent to the velum. The nearest articulatory site of the velum is the alveolar ridge; however, consonants produced there are acoustically distinct from the original velar syllables given the large distance between the articulators, which results in substantial differences in the space—and thus the resonance—within the oral cavity. Therefore, given the lack of visually and acoustically similar counterparts of the visual and auditory syllables in the non-McGurk pairings, the perceptual system struggles to reconcile them into a single, coherent percept. Instead, it typically preserves the auditory and visual syllables in their original forms or produces a combination percept that retains features from both modalities.

Given that perceiving fused percepts requires substantial visual influence, the McGurk effect has become a widely used index of the degree to which visual speech information contributes to auditory perception (Bruce & Young, 1986; Irwin et al., 2006; Ujiie & Takahashi, 2021)—a perceptual process known as audiovisual speech integration. The robustness of the

McGurk effect is demonstrated in two aspects: its irrepressible susceptibility and cross-linguistic generalizability. *First*, when presented with McGurk audiovisual pairings, people automatically perceive the integrated outcomes—even when explicitly instructed to ignore one sensory stream (Massaro, 1987; Massaro et al., 1996), remain unaware of the mismatch inputs (Burnham & Dodd, 2004), or fail to identify the original unisensory components (MacDonald & McGurk, 1978; McGurk & MacDonald, 1976)—indicating a pervasive susceptibility to forming the fusions. Neurological evidence further shows that such integration occurs in real time, without perceptual delay, comparable to the detection of an actual auditory change (Saint-Amour et al, 2007). *Second*, the McGurk effect has been consistently documented in adult native speakers of over ten languages (Bovo et al., 2009; Burnham & Dodd, 1996; Burnham & Lau, 1998; Fuster Duran, 1995; Sekiyama, 1997; Sekiyama & Tohkura, 1991, 1993; Taitelbaum-Swead & Fostick, 2016; Tiippana et al., 2010), suggesting a universality of this illusory fusion effect.

Audiovisual speech integration in prelinguistic infants

Both adults and children have been shown to demonstrate audiovisual speech integration, as indexed by the McGurk illusory effect (e.g., Hockley & Polka, 1994; McGurk & MacDonald, 1976; Sekiyama & Burnham, 2009). Worth noting, a series of seminal behavioural studies using the habituation-dishabituation paradigm suggests that even prelingual infants are already susceptible to a similar effect of visual influence in their auditory perception.

For instance, Rosenblum et al. (1997) first gaze-habituated 5-month-old English-learning infants to the audiovisual syllable /va/ (perceived as "va" due to audiovisual congruency), and then tested them with two mismatched syllable pairings in the subsequent dishabituation phase: auditory /ba/-visual /va/ (Aba-Vva, perceived as "*va*" by adults) and auditory /da/-visual /va/ (Ada-Vva, perceived as "da" by adults). Infants exhibited visual fixation recovery in response to

the Ada-Vva—but not the Aba-Vva— dishabituation pairing, suggesting that they detected a change in Ada-Vva but not Aba-Vva. This finding implies that they perceived Ada-Vva as a different sound (which should arguably be the physical auditory syllable "da"), yet Aba-Vva as the same "va" as that presented in habituation, suggesting susceptibility to a McGurk-like visual influence as that observed in adults. This interpretation was further supported by evidence that infants did not show any inherent preference for either dishabituation pairing, supporting the conclusion that infants' recovery of visual fixation reflected a perceptual change, rather than a simple attentional bias.

Desjardins and Werker (1996, 2004) complemented and extended similar findings to 4-month-old infants raised in English-language environments. Infants were habituated to either audiovisual /vi/ or /bi/ syllables (Avi-Vvi or Abi-Vbi, perceived as "vi" and "bi," respectively, due to the congruency of auditory and visual components). During test trials, all infants were then presented with the same dishabituation pairing—auditory /bi/-visual /vi/ (Abi-Vvi), which is perceived as a visually-tweaked auditory sound "vi" by adults. Infants showed no visual fixation recovery from Avi-Vvi to the dishabituation Abi-Vvi, suggesting that they, like adults, perceived the visually impacted "vi" in the Abi-Vvi pairing. In addition, a gender difference emerged: only female infants showed visual recovery when habituated to Abi-Vbi and tested with Abi-Vvi, indicating sensitivity to visual speech influence, whereas males did not. A follow-up study reversing the habituation and test pairings found the opposite pattern—only male infants showed recovery to Abi-Vbi after habituated to Abi-Vvi. These findings indicate that infants can integrate visual speech cues into auditory perception, but this integration is not reliably observed, as reflected in the inconsistent gender differences in the demonstration of novelty preference.

These studies reveal that, similar to adults, infants are influenced by visual speech information in their resulting auditory percept—for instance, visual /v/ can override auditory /b/, leading to the auditory perception of /v/. To determine whether infants can indeed perceive a third, fused percept that deviates from both the auditory and visual speech components, Burnham & Dodd (2004) tested 4.5-month-olds on the McGurk effect using the classic pairing of auditory /ba/-visual /ga/ (Aba-Vga), which is perceived as "da" if the illusion occurs. Prior to auditory-only test trials featuring /ba/ (the original auditory stimulus), along with /da/ and /tha/ (two possible fused percepts), infants in the experimental and control groups were habituated to the McGurk pairing (Aba-Vga) and a congruent pairing (Aba-Vba), respectively. The results supported the conclusion that infants perceive speech in an audiovisual manner, as infants in the experimental group recognized /da/ and /tha/ as more familiar than /ba/, whereas the control-group infants showed no such preference. Kushnerenko et al. (2008) further corroborated this early integration capacity using event-related brain potentials (ERPs). They found that 5-month-olds exhibited no audiovisual mismatch response (AVMMR) when presented with Aba-Vga pairings, suggesting successful assimilation into a unified percept. In contrast, when exposed to conflicting cue combinations (Aga-Vba), infants did show a mismatch response, indicating failure to integrate the audiovisual inputs into a single coherent sound.

Evidence against a robust early audiovisual speech integration

Despite the abovementioned evidence supporting infants' early capacity to integrate cross-modal speech information—as indexed by their early susceptibility to the McGurk illusion effect—two lines of research suggest that such ability may not be robust during infancy. *First,* as discussed, though some studies using the McGurk effect paradigm report that young infants already demonstrated an adult-like fusion effect (Burnham & Dodd, 2004; Rosenblum et al.,

1997), this capacity might be mandatory and not as reliable (Desjardins & Werker, 2004). Indeed, substantial evidence suggests that the McGurk effect continues to develop well beyond infancy. *Second*, infants fail to show signs of audiovisual speech integration when exposed to a racially unfamiliar speaker, suggesting that their McGurk effect is subjected to visual disruptions.

In this section, we first reviewed the protracted developmental trajectory of the McGurk effect beyond infancy and consider potential explanations. We then looked at the interference on early audiovisual speech integration imposed by other-race faces. Together, these two lines of evidence provide the theoretical basis for our own research questions.

A protracted developmental trajectory of AV-speech integration

As a foundational prerequisite for verbal communication, audiovisual speech integration has been demonstrated to experience a developmental increase throughout childhood, and possibly even beyond. In their seminal McGurk effect study, McGurk and MacDonald (1976) presented preschoolers (3–4 years), elementary school children (7–8 years), and adults (18–40 years) with four pairs of mismatched auditory and visual syllable pairs, including two McGurk stimuli (Aba-Vga and Apa-Vka) and two non-McGurk stimuli (Aga-Vba and Aka-Vpa)—created by reversing the auditory and visual components of the McGurk pairs—and measured their perceptual outcomes through self-report. Older participants were found to be more subjected to the influence from visual cues than younger participants, with adults, school children, and preschoolers showing 92%, 52%, and 59% visually impacted responses, respectively; the two younger groups showed comparable performance. Worth noting, this developmental increase was not related to differential auditory processing, as all age groups exhibited high accuracy in the auditory-only condition (91%, 97%, and 99% for preschoolers, school-age children, and adults), where they were asked to repeat what they heard in the absence of visual cues. Interestingly, both

younger groups frequently reported the fused precepts when exposed to the two McGurk pairs, suggesting an existing—yet still-developing—audiovisual speech integration mechanism.

This gradual increase in visual speech influence on audiovisual speech perception has been confirmed in a series of subsequent studies. Massaro (1984) compared this integration in 5–7-year-olds and adults. Five synthetic auditory syllables, ranging along a /ba/-/da/ continuum, were factorially combined with three visual articulation conditions (i.e., /ba/, /ga/, and no articulation) to assess participants' speech identification. Children demonstrated about only half the visual influence observed in adults (visual effect sizes: 33% for children and 75% for adults). As in previous findings, this quantitative difference was not attributable to disparities in auditory processing capacities, as perceptual sensitivity to the auditory components in the absence of visual cues (i.e., in the no-articulation conditions) was comparable across children and adults. Using a similar factorial combination design, Massaro et al. (1986) reported consistent findings that adults showed a stronger visual influence than 4–6-year-olds, who reported 82% and 35% visually influenced responses, respectively. Further studies have shown that although the degree of visual influence is low in early childhood, it increases progressively around ages 6–8 (Sekiyama & Burnham, 2008), begins to approximate adult-like patterns by ages 10–12, and continues developing beyond the age of 12 (Hockley & Polka, 1994; Vannasing et al., 2024).

The aforementioned studies suggest that while the perceptual strategy of fusing auditory and visual information in speech perception remains stable from early childhood to adulthood, *cue weighting* changes over time: auditory cues dominate bimodal speech perception in early childhood, whereas visual cues become increasingly influential with age. Three hypotheses have been proposed to explain this age-related shift. First, the *attentional hypothesis* posits that children pay less attention to visual speech cues than adults, thereby reducing their impact (see

Fuzzy Logical Model of Perception [FLMP] in Massaro, 1984 for potential counterevidence). Second, the *perceptual hypothesis* argues that lip-reading accuracy, which has been tested to strongly and positively associated with the extent of visual influence on bimodal speech perception, was poorer in children than adults (Massaro et al., 1986). Lastly, the *articulatory hypothesis* proposes that the enhanced visual influence is attributable to an accumulating speech production experience (Desjardins et al., 1997; Siva et al., 1995). The first two reflect perceptual tuning accounts, while the third is rooted in sensorimotor development. If these hypotheses hold, audiovisual speech integration may be even more limited in prelinguistic infants.

Building on these accounts of a chronologically enhancing integration, we propose that the integration capacity is likely still limited but gradually strengthens across the second half of the first year of life. This period marks a developmental intersection: their attentional system undergoes rapid maturation, and speech perception becomes increasingly shaped by infants' speech production. From 6 months on, infants develop endogenous attention, enabling them to voluntarily direct their gaze toward socially and linguistically relevant features—such as articulating mouths—to support speech processing (Colombo, 2001; Richards et al., 2010). Meanwhile, infants begin to produce canonical babblings (Oller, 2000), a crucial developmental milestone that may foster emerging metalinguistic awareness (Bergelson, 2020) and increase their sensitivity to visual articulatory cues (Chandrasekaran et al., 2009). Such dual development—of attentional control and vocalization practices—likely contributes to a more flexible and functionally-driven attention to visual speech cues, facilitating their integration with auditory speech perception. Indeed, several studies have shown that infants increasingly fixate on speakers' lips during this period (Hunnius & Geuze, 2004; Lewkowicz & Hansen-Tift, 2011), suggesting an active incorporation of visual articulatory information. Thus, as speech perception,

production, and attention become increasingly coordinated, we hypothesize that audiovisual speech integration follows a developmental trajectory that strengthens throughout the latter half of the first year.

The other-race effect (ORE) on infants' audiovisual speech integration

Another thread of evidence against a resilient audiovisual speech integration in infancy is the other-race effect (ORE). The ORE in the unimodal race-based face processing domain refers to the phenomenon whereby individuals more accurately recognize faces of their own race compared to faces of other racial groups, suggesting a perceptual advantage for own-race faces within the human face processing system (Bothwell et al., 1989; MacLin & Malpass, 2001; Valentine, 1991). This effect is thought to arise from greater exposure to—and more frequent interactions with—own-race faces beginning early in development (Anzures et al., 2012; Bar-Haim et al., 2006). Indeed, an extensive analysis of video-recordings captured from infants' first-person point of view revealed that own-race faces dominated nearly 96% of infants' visual exposure to all faces from an early age (Sugden et al., 2013).

This high statistical distribution of own-race faces from early in life may account for the early emergence and clear manifestation of the ORE in infancy. For instance, Kelly et al. (2007) examined face discrimination abilities in 3- to 9-month-old Caucasian infants who had predominantly experienced Caucasian faces since birth. They found that at 3 months of age, infants could discriminate faces across multiple racial groups (e.g., African, Chinese, Middle Eastern), but by 6 months, their sensitivity narrowed, with successful discrimination observed only for Caucasian and Chinese faces. By 9 months, infants showed discrimination exclusively for own-race (Caucasian) faces. This developmental narrowing has been replicated in Asian infants as well (Kelly et al., 2009). Collectively, these findings suggest that ORE in face recognition emerges by 6 months of age and becomes fully established by 9 months.

ORE is argued to stem from the *perceptual narrowing* (or *perceptual tuning*) process, where the initially broad and sensitive perceptual abilities become increasingly pruned by and attuned to the most prevalent information in infants' immediate surroundings (e.g., Maurer & Werker, 2014; Pascalis et al., 2002). Beyond extensively studied unimodal perceptual narrowing—such as on visual processing of faces from different races—recent scholarly discussions explore whether this narrowing process is modality-general, or pan-sensory, affecting how multiple sensory inputs are integrated (Lewkowicz & Ghazanfar, 2009; Pons et al., 2009).

This possibility has been directly tested in a series of studies, which investigated whether the perceptual attunement to own-race faces in the latter half of the first year in infancy influences the integration of facial and auditory speech cues. Japanese infants aged 5–6 months and 8–9 months were familiarized with a McGurk stimulus (Apa-Vka, perceived as /ta/ if the audiovisual integration occurs) enacted by own-race (East-Asian) and other-race (Caucasian) faces. In the subsequent testing phase, infants who previously viewed own-race faces showed a novelty response to the auditory /pa/ sound, suggesting that they had perceived the fused /ta/ during familiarization (Ujiie et al., 2021, 2020). In contrast, infants familiarized with the same McGurk stimulus enacted by other-race faces did not show a significant preference for the auditory /pa/, indicating a lack of integration when exposed to unfamiliar-race faces. Supporting this behavioural evidence, functional near-infrared spectroscopy (fNIRS) data showed that 8–9-month-olds who viewed own-race faces during McGurk stimulus presentation (Apa-Vka) exhibited activation in the left temporal region, a key area for audiovisual speech processing (e.g., Beauchamp et al., 2004; Calvert et al., 2000). However, no significant activations were observed in infants who viewed other-race faces (Ujiie et al., 2020). This converging evidence

revealed that infants demonstrate audiovisual speech integration when viewing own- but not other-race faces, suggesting an ORE in infants' bimodal speech perception.

Despite the convincing neurological evidence suggesting that other-race faces can impede the integration of audiovisual speech cues in infancy, we question the experimental paradigms used to generate the behavioural evidence for the ORE in these studies, particularly from the standpoints of recognition memory and visual attention. Specifically, Liu et al. (2011) found that infants exhibited reduced recognition memory for other-race faces, which was associated with decreased visual attention to the internal facial features of these racially unfamiliar faces. Therefore, the familiarization paradigm might not be the most ideal design for measuring the potential ORE in infants' audiovisual speech integration. More specifically, it cannot rule out the possibility that infants were able to perceive the McGurk fusion in real time during the familiarization phase but failed to retain the fused auditory percept well into the test stage, thereby preventing a novelty preference from emerging (Ujiie et al., 2021, 2020)—possibly due to memory-related constraints.

In sum, this section reviewed two lines of evidence suggesting that audiovisual speech integration may remain limited in infancy. First, studies employing the McGurk effect indicate that the bimodal speech perception showed an age-related strengthening that extends beyond infancy into childhood, potentially due to experience-driven attentional, speech perception, and speech production mechanisms; at the same time, these possibilities suggest that infants may show improving integration across the second half of the first year, as emerging endogenous attention and the intersection of perceptual and vocal capacities are likely to facilitate infants' visual exploration of articulatory speech cues. In addition, unfamiliar-race faces appear to disrupt the integration process, hindering the emergence of the McGurk effect in the latter half of the

first year (i.e., ORE). However, because prior studies rely on memory-based experimental paradigms, the underlying mechanisms of this interfering ORE remains insufficiently understood.

The present study

The current study comprises a set of three experiments designed to answer two main research questions. We first asked whether infants demonstrate audiovisual speech integration and, if so, whether this capacity strengthens across infancy. We then examined whether faces of other races interfere with infants' ability to integrate auditory and visual speech information and, if so, what could tentatively account for this disruption.

To address these two questions, we sampled White infants aged six to 12 months for three reasons. First, as discussed earlier, prior research indicates that infants start to show the McGurk effect before 6 months (e.g., Burnham & Dodd, 2004; Kushnerenko et al., 2008), providing a theoretical starting point. Second, from 6 months onward, the increasing interaction between speech perception and production—as well as the emergence of endogenous attention—may reinforce infants' attention to the visual source of auditory information, potentially facilitating age-related enhancements in audiovisual speech integration (i.e., the 'differentiation' account for progressive multisensory development from a Gibsonian perspective). Finally, the other-race effect in facial processing has been shown to emerge during the second half of the first year in infancy, potentially interfering with the integration of auditory and facial speech cues (i.e., the perceptual tuning account for regressive multisensory development). Thus, if an other-race effect on audiovisual speech integration exists, it is most likely to surface within this developmental window, as previous literature might have suggested.

In Experiment 1, we examined infants' susceptibility to the McGurk effect using faces of their own races (i.e., White faces), hypothesizing that audiovisual speech integration is already present but still developing within the tested age range. Experiment 2 presented infants with faces of an unfamiliar race (i.e., East Asian faces) to investigate the potential other-race effect (ORE) on audiovisual speech integration. We predicted that an ORE would emerge during this period, resulting in a weakened McGurk effect that would show no—if not a declining—age-related trend due to the perceptual narrowing in processing other-race faces (e.g., Kelly et al., 2007, 2009), which constitutes one of the two sensory modalities necessary for the McGurk illusion to occur. If an ORE in audiovisual speech perception were observed, we hypothesized that it would be attributable to differential face-scanning strategies employed by infants when processing own- versus other-race faces, particularly distinct attention allocated to the mouth. Experiment 3 was designed to provide theoretical support for our interpretation of audiovisual speech integration by exposing infants to alternating and non-alternating syllables that were always congruent across auditory and visual modalities (see behavioural task below).

To measure whether infants integrate auditory and visual information during speech perception, we designed a behavioural task, in which infants were presented with a sequence of *identical* auditory syllables dubbed with visual displays of faces. Crucially, on every other auditory presentation (e.g., auditory /ba/), the faces produced articulations that could trigger the McGurk illusion (e.g., visual /ga/)—potentially leading to a perceptual shift between the actual auditory input (auditory /ba/) and an illusory, fused percept (auditory /da/). If infants were indeed integrating audiovisual speech cues during these articulating moments, they would experience a regular alternation between two distinct phonetic percepts (i.e., /ba-da-ba-da/). Conversely, in a control condition lacking the McGurk components (e.g., repeating auditory /ga/ paired with

visual /ba/ on every other auditory syllable), the auditory input should be perceived as a constant repetition of the same auditory syllable (i.e., /ga-ga-ga-ga/). We indexed infants' auditory integration by comparing their looking times across the two conditions, with particular interest in whether they displayed a visual preference for either audiovisual speech condition, building on prior findings that infants prefer sequences featuring alternation over repetition (Kidd et al., 2012). While conceptually grounded in the Stimulus-Alternation Preference Procedure (SAPP; Best & Jones, 1998), our method introduces a key innovation: the perceived alternation in our task arises from an illusory fusion effect rather than from physically different auditory stimuli.

In contrast to the conventional habituation or familiarization paradigms commonly used to investigate the McGurk effect in infancy, the current approach offers notable improvements in both measurement sensitivity and ecological validity. First, it minimizes reliance on infants' memory, providing a more contingent examination of perceptual capacity rather than memory for specific syllables. Second, it assesses changes in auditory perception without making predefined expectations about the exact illusory percept. In other words, it allows any perceptual change elicited by visual articulations to be detected, as opposed to previous methods that often assumed a particular outcome (e.g., hearing a /da/ in the Aba-Vga pairing). Finally, the short trial duration (20s) and total trial numbers (16 trials) enables the inclusion of a variety of syllables and facial stimuli, boosting the generalizability of results and helps mitigate potential biases arising from stimulus-specific effects—a frequent constraint in studies using a narrow range of exemplars.

Experiment 1

Participants.

Thirty-three full-term English-learning White Canadian infants (13 females) with normal vision and hearing participated in the current experiment after caregivers provided informed

consent. The participants were from 190 to 350 days old (M = 8.65 months, SD = 1.47 months) and were recruited from the Southern Ontario area in Canada. According to parental reports, all participating infants primarily interacted with White individuals and were exposed to English in their everyday life at the time of participation.

Twenty-four additional infants participated but were excluded from data analysis because of failure to complete the experimental procedure due to fussiness (n = 7), calibration failure (n = 5), equipment glitch (n = 1), primary caregivers' racial category being non-White (n = 7)[2], continuous parental interference throughout the study session (n = 1), or performance that was more than two standard deviations from the group mean (n = 3).

All experimental protocols, including the procedures for obtaining caregivers' informed consent, were reviewed and approved by the Research Ethnics Board of McMaster University (approval no.: 3665). The participating families received a book, tote bag, or T-shirt of their choice in appreciation for their participation.

Stimuli.

We video recorded one young female articulating four syllables, including /ba/, /pa/, /ga/, and /ka/, with a neutral facial expression. Each syllable was articulated for approximately 800 ms. We trimmed the recording of each syllable into 1 second clips and further edited the resultant audio and visual components (referred to as auditory syllables and visual syllables henceforth) of the recording to create the stimuli for the current study.

For the ***auditory syllables***, we used Adobe Audition to remove the background noises from the original recordings and matched the overall loudness across the four syllables. For the

[2] These seven non-White primary caregivers self-identified as follows: Black (n = 1), biracial (n = 2; Chinese and Pakistani, Black and White), Hispanic (n = 1), Vietnamese (n = 1), Salvadoran (n = 1), and Turkish (n = 1). The rationale for their exclusion was to ensure that the experimental condition represented a strictly own-race face condition.

visual syllables, we first used a generative adversarial network (GAN) to copy the facial movements of the young female to 16 new faces (eight White and eight East-Asian). This process generates videos of photo-realistic faces that make identical facial movements to those in the original recording, thereby greatly reducing the possibility that our findings are biased by idiosyncratic facial movements across individuals—in other words, minimizing the stimulus-dependent confounds introduced by different faces (Tiippana et al., 2023). Next, we digitally removed the hair from each of the 16 faces, standardized their sizes, and placed them against a light gray background. All other visual properties of these faces were unedited to preserve the natural appearance of each face, thereby maximizing the ecological validity of the visual stimuli.

To examine infants' audiovisual integration, we created two types of audiovisual pairings by combining the videos of the visual syllables and sounds of the auditory syllables introduced above. As shown in Table 1, the ***McGurk pairs*** were audiovisual combinations that were likely to induce the perception of the McGurk illusion. For instance, an illusory auditory perception of /da/ would ideally arise when the visual syllable /ga/ is paired with the auditory syllable /ba/; similarly, when the visual syllable /ka/ is dubbed with the auditory syllable /pa/, a /ta/ percept would emerge (e.g., McGurk & MacDonald, 1976; Tiippana et al., 2023). The ***non-McGurk pairs***, on the other hand, were audiovisual combinations that were unlikely to lead to the McGurk effect. For example, instead of having visual /ga/-auditory /ba/ as in the McGurk pair, the non-McGurk pair composed of visual /ba/-auditory /ga/. In this case, the perceptual outcome would be highly probable to remain being the physical auditory syllable, /ga/, which has been corroborated previously in children using behavioral self-report (McGurk & MacDonald, 1976) as well as inferred in infants using electrophysiological measurement (Kushnerenko et al., 2008). As explained in *Introduction*, the rationale for the absence of illusion in the non-McGurk pairs

resides in the phonotactic constraints of the stimulus phonemes. The articulatory mechanism for bilabial consonants (i.e., sounds produced by bringing both lips together), such as the visual syllables /ba/ and /pa/ in the non-McGurk pairs, naturally constrains the auditory perceptual outcomes to syllables starting with a bilabial consonant (i.e., /b/, /p/, or /m/). However, the actual auditory syllables used in the non-McGurk pairs—/ga/ and /ka/—both start with velar consonants, which are articulated at the back of the oral cavity and do not involve visible lip contact. The mismatch in the places of articulation makes /ga/ and /ka/ both phonetically and visibly distinguishable from /ba/ and /pa/. As a result, the non-McGurk pairs are impossible to elicit the fused (or *altered*) auditory percepts in the same way as the McGurk pairs.

Worth noticing, in the current study, we created the non-McGurk pairs by reversing the visual and auditory syllables in the McGurk pairs. This design has two advantages: 1) it avoids the possibility that infants' differential responses to the two types of audiovisual combinations (i.e., McGurk vs. non-McGurk pairs) were driven by the mismatch between the auditory and visual information, as both adults (Green & Kuhl, 1991) and infants (Burnham & Dodd, 2004) have been found to demonstrate longer looking times to mismatched compared to matched auditory-visual stimuli. In this design, the auditory and visual stimuli were mismatched in both pair types, controlling for this potential confound; 2) the two pair types presented the same auditory and visual information, eliminating the possibility that infants' response being driven by certain syllables presented only in one condition. As will be introduced in the following *Procedure* section, we used the McGurk and non-McGurk pairs to create the McGurk and non-McGurk trials, respectively.

Table 1. The auditory and visual syllable combinations of the McGurk and non-McGurk pairs used in the current study.

	Visual syllables	Auditory syllables	Theoretical percepts
McGurk pairs	/ga/	/ba/	altered (/da/)
	/ka/	/pa/	altered (/ta/)
non-McGurk pairs	/ba/	/ga/	unaltered (/ga/)
	/pa/	/ka/	unaltered (/ka/)

Procedure.

Throughout the study, infant participants watched videos of faces articulating visual syllables and listened to the acoustic sounds of auditory syllables across a maximum of 16 trials. In each ***McGurk trial***, infants watched one McGurk pair (1 second) repeating 20 times. For every alternating iteration (i.e., the 2nd, 4th, 6th, 8th, 10th, 12th, 14th, 16th, 18th, and 20th iteration), the video will be held still from the first frame, creating the presentation of a still face image paired with the audio syllable sound. In other words, infants would see the regular alternations of an articulating face and a still-face image while hearing the same syllable sound repeating twenty times throughout the trial. Should infants be able to reliably integrate auditory and visual speech information, we expected their auditory perception to be altered only when the face was articulating (i.e., the 1st, 3rd, 5th, 7th, 9th, 11th, 13th, 15th, 17th, and 19th iteration) but not when it remains still (i.e., the 2nd, 4th, 6th, 8th, 10th, 12th, 14th, 16th, 18th, and 20th iteration). Each ***non-McGurk trial*** shared an identical structure with the McGurk trial, with the only difference being that the non-McGurk pairs were presented. Because non-McGurk pairs were unlikely to alter auditory perception, the infants should hear the same auditory syllable repeating twenty times

throughout the non-McGurk trials. Figure 1 illustrates the visual and auditory syllable pairings along with the ideal perceptual auditory outcomes in the McGurk and non-McGurk trials.

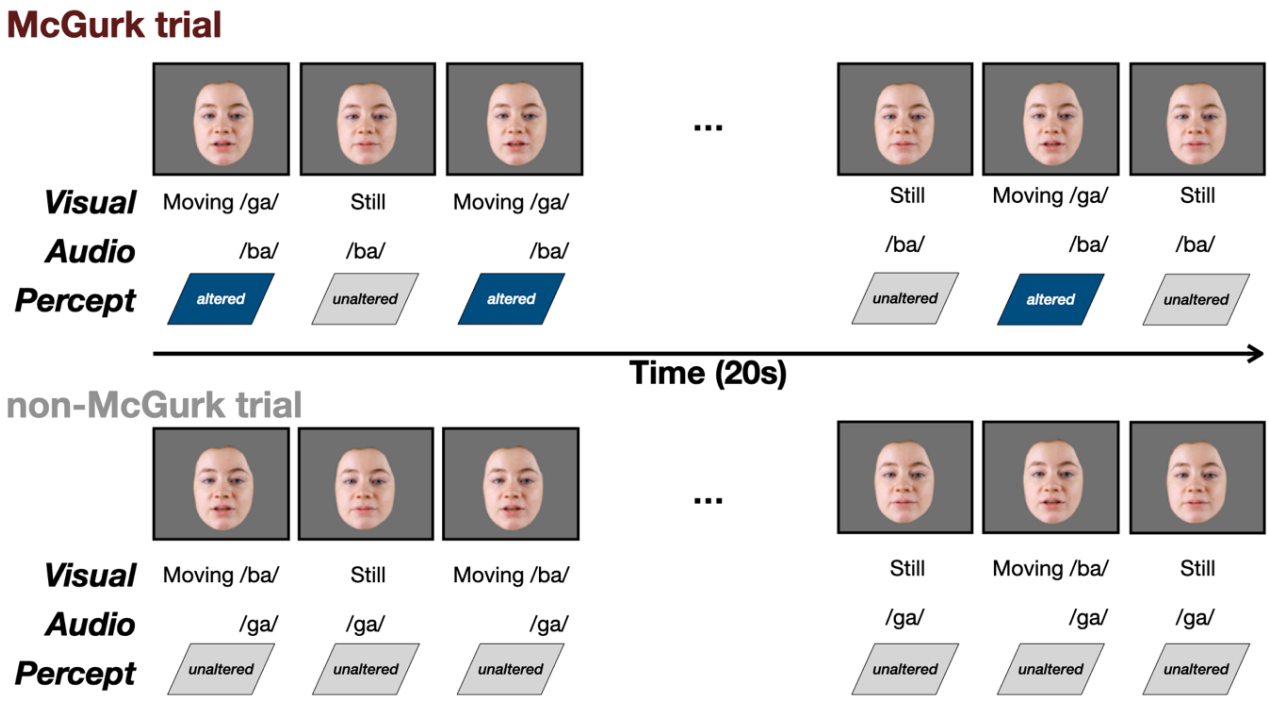


Figure 1. *Schematic presentation of the experimental procedures in the McGurk and non-McGurk trials (Experiment 1; own-race face condition).* One among the eight White stimuli faces and two among the four audiovisual syllable pairs (i.e., visual /ga/-auditory /ba/ and visual /ba/-auditory /ga/) are presented here for demonstration.

This alternating presentation allowed us to assess infants' audiovisual integration by comparing their total looking times between the McGurk and non-McGurk trials. Historically, researchers have measured the McGurk illusion by examining how consistently individuals perceive the illusory percept, noting that some individuals experience it consistently whereas others do so only sporadically (e.g., Sekiyama & Burnham, 2008). In a similar vein, we anticipate that infants' audiovisual integration capacities will not be strictly dichotomous (i.e.,

either ‘present’ or ‘absent’), but rather, will vary in the consistency with which they engage in the integration process. These varying capacities should produce three types of distinct looking patterns, as illustrated in Figure 2 below.

First, if infants have developed a robust ability to integrate visual articulatory information into their auditory perception, they would perceive a *regular alternation between two auditory sounds*—likely the illusory percept and the physical auditory syllable—in the McGurk trials (see Figure 2a), as opposed to a repeated auditory syllable in the non-McGurk trials (see Figure 2c). Because the alternating sound sequence in the McGurk trials is more patterned than the repetitive sequence in the non-McGurk trials (Al Roumi et al., 2023) and infants tend to prefer structured patterns if they are not perceptually overcomplicated (Kidd et al., 2012, 2014), we predicted longer looking times in the McGurk trials than in the non-McGurk trials.

Second, infants’ audiovisual integration might still be under development. Under this circumstance, the integration of visual and auditory speech information could occur only sporadically, leading to an *irregular mix of altered and unaltered sounds* (see Figure 2b) in the McGurk trials. Since infants generally show reduced interest in unpredictable sequences (Kidd et al., 2012, 2014), we anticipated that this unstable integration developmental stage would yield longer looking times in the non-McGurk than in the McGurk trials.

Lastly, if infants are unable to incorporate the visual speech information into auditory perception at all, their perceptual outcomes should faithfully reflect the *repetitive sounds* of the physical auditory syllable. In other words, they would hear the same sound repeating throughout both the McGurk and non-McGurk trials, thus showing comparable looking times across the two trial types.

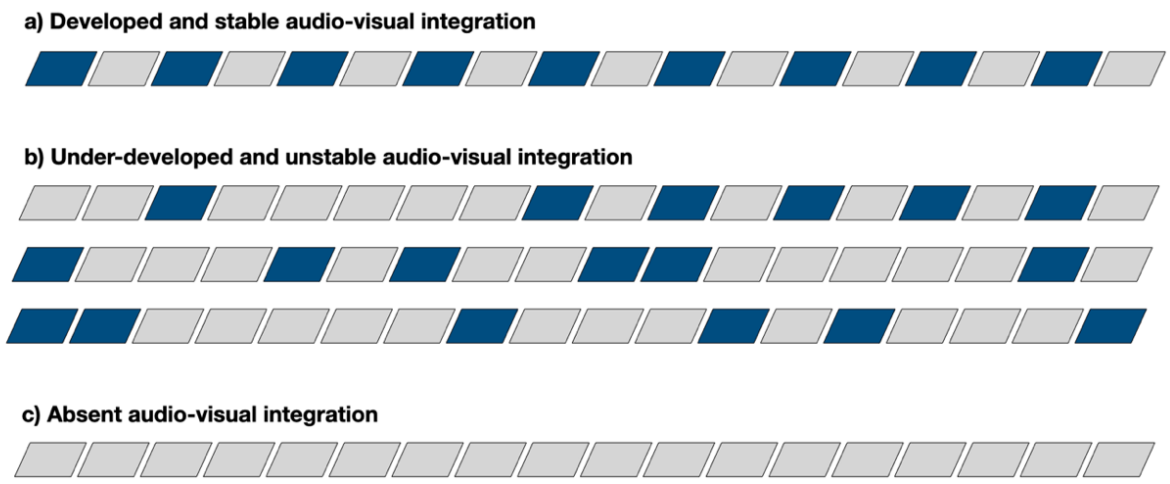


Figure 2. *Three possible auditory perceptual outcomes in the McGurk trials.* The blue and grey blocks represent altered and unaltered auditory perception, respectively. a) An alternating and consistent auditory perception sequence derived from a fully developed audiovisual integration capacity. b) An unpredictable auditory sequence resulting from an under-developed audiovisual integration ability. c) A repetitious auditory perception in the absence of audiovisual integration.

To determine if infants could integrate audiovisual information, we referred to their on-screen looking time measured by an EyeLink 1000 Plus eye-tracker (500 Hz sampling rate, SR Research, Canada) as a proxy. It is important to note that we anticipated infants' looking time to closely reflect their auditory perception. However, since infants are sensitive to moving stimuli, it is likely that the moving face on the screen captures their visual attention even though they were already bored with the sequence of the auditory. This may lead to longer looking times that do not necessarily indicate sustained interest. To ensure that infants' looking behaviours were guided by their auditory perception rather than dynamic visual properties, we implemented a gaze-contingent design. Specifically, the face would move only when infants were looking at it but would turn semi-transparent and remain still when infants looked away from the screen. A

similar design has been used in previous studies on infants' perception involving moving stimuli (e.g., Xiao et al., 2023).

The entire experimental procedure included four blocks in total, with four trials within each block: two McGurk trials (visual /ga/-auditory /ba/ Aba-Vga; visual /ka/-auditory /pa/ Apa-Vka) and two non-McGurk trials (visual /ba/-auditory /ga/ Aga-Vba; visual /pa/-auditory /ka/ Aka-Vpa). The blocks were played to the infants in a randomized order, and no identical trial was presented twice in a row. The experiment session terminated automatically after participants finished four blocks of four trials each (totaling 16 trials) or if they stopped looking to the screen.

Infants sat on their caregivers' laps in a sound-attenuated testing room throughout the experiment. Prior to the experiment, the caregivers were instructed not to interfere with their infants' looking behaviours, such as by directing infants' attention to the screen by pointing or verbal prompts. Infants sat approximately 60 to 80 cm away from the 55 × 31 cm^2 (25-inches) computer monitor, where we used Psychtoolbox (3.19.8) software to play the visual stimuli. The experiment started with an infant-controlled calibration program to ensure eye tracking precision and accuracy. During calibration, a cartoon figure was presented on the screen. Immediately after infants fixated on the animation target, it would move to another position with a rewarding rattling sound. The calibration procedure was completed once infants successfully fixated at five locations (four corners and the center). At the beginning of each trial, an attention-getter (i.e., a colorful bouncing circle) located at the center of the monitor directed infants' attention (back) to the screen.

Results and Discussion.

We first filtered the raw eye-tracking data to calculate fixation data, which was generated according to the default setting in EyeLink's DataViewer software (ver. 4.4.1). We removed trials in which participants' total fixation time was less than 500 ms, excluded the last block if parents

began to interfere infants' looking behaviours toward the end of the study session[3], and averaged each participant's looking time in McGurk and non-McGurk trials to index their perception. We measured the McGurk effect by examining whether infants showed a significant looking preference for one type of trial over the other. Therefore, we interpreted infants' longer looking time to either trial type as evidence of audiovisual integration.

On average, infants finished 15.4 trials. For the following analyses, we collapsed looking time data across trials of the same type (i.e., McGurk or non-McGurk), even though each trial type included two different audiovisual syllable pairings. This decision was based on the control analyses showing that infants looked comparably to the /BaGa/ (Aba-Vga and Aga-Vba) and /PaKa/ (Apa-Vka and Aka-Vpa) pairings. These findings (see Syllable-specific Control Analysis in the *Supplementary Materials* for details) suggest that specific syllables did not systematically influence infants' looking behaviour. In addition, infant gender did not affect infants' looking preference across trial types (see Gender Control Analysis in the *Supplementary Materials* for details). Therefore, data were combined across infant gender for the following analyses.

Infants showed evidence of audiovisual integration.

To determine if infants exhibited audiovisual speech integration, we conducted a paired-sample *t*-test comparing their average total looking times to the McGurk and non-McGurk trials. The analysis revealed that infants looked significantly longer at McGurk (M = 11.60 s, SD = 2.83) than at non-McGurk trials (M = 11.06 s, SD = 3.30; t (32) = 2.41, p = .022; Cohen's d = 0.42; see Figure 3). This suggests that infants processed the two types of trials differentially,

[3] Two infants completed all four blocks (i.e., 16 trials), but their caregivers began to interfere during the final block as infants stopped looking to the screen. Thus, looking data from the last block was excluded for these two infants.

which we interpret as evidence that White Canadian infants exhibited audiovisual speech integration when viewing own-race faces during the tested age range (6 to 12 months).

Prior to the paired-sample *t*-test, we confirmed that the looking time differences between the two trial types—calculated by subtracting each infant's total looking time in the non-McGurk trials from that in the McGurk trials—were normally distributed, as indicated by a Shapiro–Wilk test ($W = 0.98$, $p = .773$) and visual inspections of the histogram and Q-Q plot. These results supported the normality assumption, further justifying the appropriateness of the analysis.

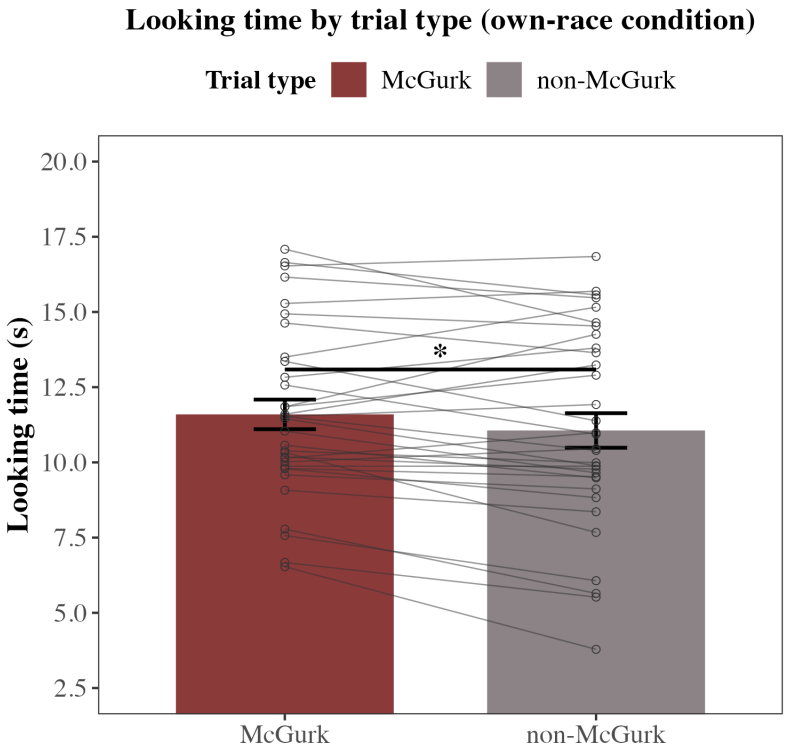


Figure 3. *Mean looking times to the McGurk and non-McGurk trials in own-race face condition (Experiment 1).* The asterisks represent the statistically significant difference in the looking times between two types of trials (* $p < .05$). Error bars represent ±1 standard error from the mean. Each dot represents an individual participant's looking time in each trial type, and the lines connecting paired dots reflect within-participant comparisons.

Audiovisual speech integration was stable across infancy.

To examine whether infants' audiovisual integration showed any developmental change, we performed a Pearson correlation between infants' age measured in days and their McGurk preference scores (i.e., the difference in average total looking times between the two types of trials), calculated as:

$$\text{McGurk preference score} = \text{Total looking time}_{\text{McGurk}} - \text{Total looking time}_{\text{non-McGurk}}$$

The correlation was not significant ($r(31) = -.15$, $p = .415$; Figure 4), suggesting that audiovisual speech integration remained stable in White Canadian infants between 6 and 12 months of age. This finding further implies that the capacity for integration likely emerges earlier than 6 months, consistent with previous findings (Burnham & Dodd, 2004; Rosenblum et al., 1997).

Before the correlation analysis, we ensured that the assumptions of linearity and homoscedasticity were met, and the residuals from the linear model (*diff ~ ageDays*) were normally distributed ($W = 0.97$, $p = .568$). Cook's distance analysis identified one potentially influential observation. However, excluding this observation did not substantially alter the results or the overall interpretation: the correlation increased slightly ($r(30) = -.31$) but remained statistically non-significant ($p = .087$). Therefore, for the purpose of consistency, the full dataset was retained in all reported analyses.

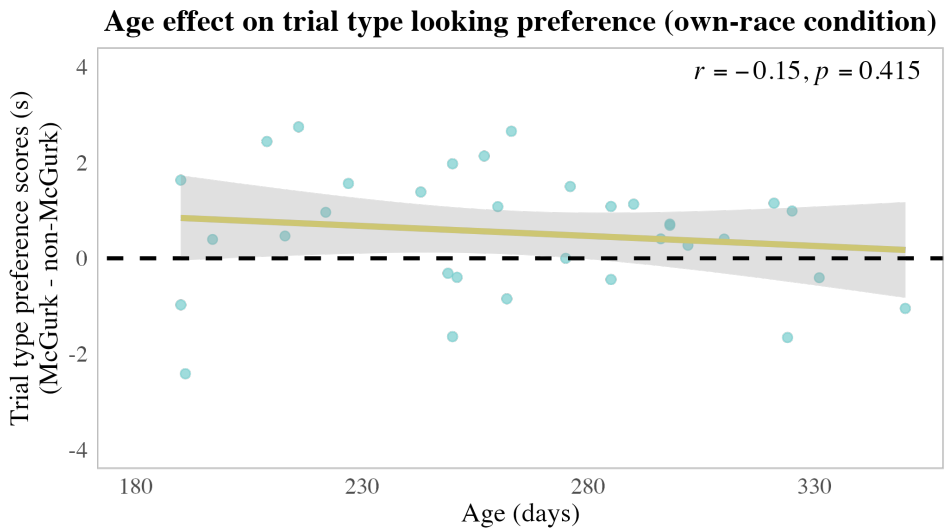


Figure 4. *Age-related change in the trial looking preference in own-race face condition (Experiment 1).* Each dot represents data from an individual participant. The horizontal dashed line indicates equal looking to McGurk and non-McGurk trials. Dots above and below the dashed line are individuals who looked longer to the McGurk and non-McGurk trials, respectively. The shaded area around the regression line represents the confidence interval of the regression model.

Based on the three possible auditory perceptual outcomes and their corresponding levels of audiovisual speech integration outlined in the *Procedure* section (see Figure 2), we inferred that White Canadian infants' significantly and consistently longer looking times toward the McGurk trials may reflect a fully-fledged capacity to integrate auditory and visual speech information throughout the latter half of the first year. This finding supports our first hypothesis that speech-related audiovisual integration ability is already present in infancy but diverges from our predication that this capacity would show an age-related strengthening across infancy. Instead, the data showed that this capacity may already be robust by the time infants reach six months of age (i.e., the time of their participation in this study).

In Experiment 2, we continue to explore White Canadian infants' audiovisual speech integration, focusing on its resistance to faces of an unfamiliar race as a visually distracting factor. As prior research has suggested (see also *Introduction*), underrepresented faces in infants' perceptual environments (e.g., other-race faces) can significantly impair their ability to engage in the McGurk illusory effect (e.g., Ujiie et al., 2020, 2021). However, in addition to the unideal use of familiarization paradigm to test other-race effect (ORE) on audiovisual speech integration, that study exclusively tested East-Asian infants born and raised in Japan, a racially homogenous society where infants receive negligible visual exposure to faces outside their own race. Thus, it remains unclear whether Western infants—particularly those growing up in the racially diverse environment such as Southern Ontario, where our participants were recruited—would exhibit similar susceptibility to the distracting effects of other-race faces. To address this question, we recruited a new cohort of White Canadian infants and assessed their audiovisual integration using other-race (i.e., East-Asian) faces in Experiment 2.

Experiment 2

Participants.

Twenty-one full-term English-learning White infants (10 females) with normal vision and hearing participated after caregivers provided informed consent. The participants were from 196 to 341 days old (M = 8.93 months, SD = 1.41 months) and were recruited from the Southern Ontario area in Canada. According to parental reports, all participating infants primarily interacted with White individuals, were exposed to English in their daily lives, and had minimal exposure to East-Asian faces at the time of participation.

Additional fourteen infants' data were excluded from data analyses due to fussiness (n = 2), continuous parental interference throughout the study session (n = 2), the primary caregiver's

racial category being non-White (n = 9)[4], and performance that was more than two standard deviations from the mean (n = 1). All experimental protocols, including the procedures for obtaining the informed consent, were reviewed and approved by the Research Ethnics Board of McMaster University (approval no.: 3665). The participating families received a book, tote bag, or T-shirt of their choice in appreciation for their participation.

Stimuli & Procedure.

The experimental stimuli and procedures were identical with those used in Experiment 1 except that the infant participants watched East-Asian female faces instead of White female faces. See Figure 5 for schematic representations of the procedures.

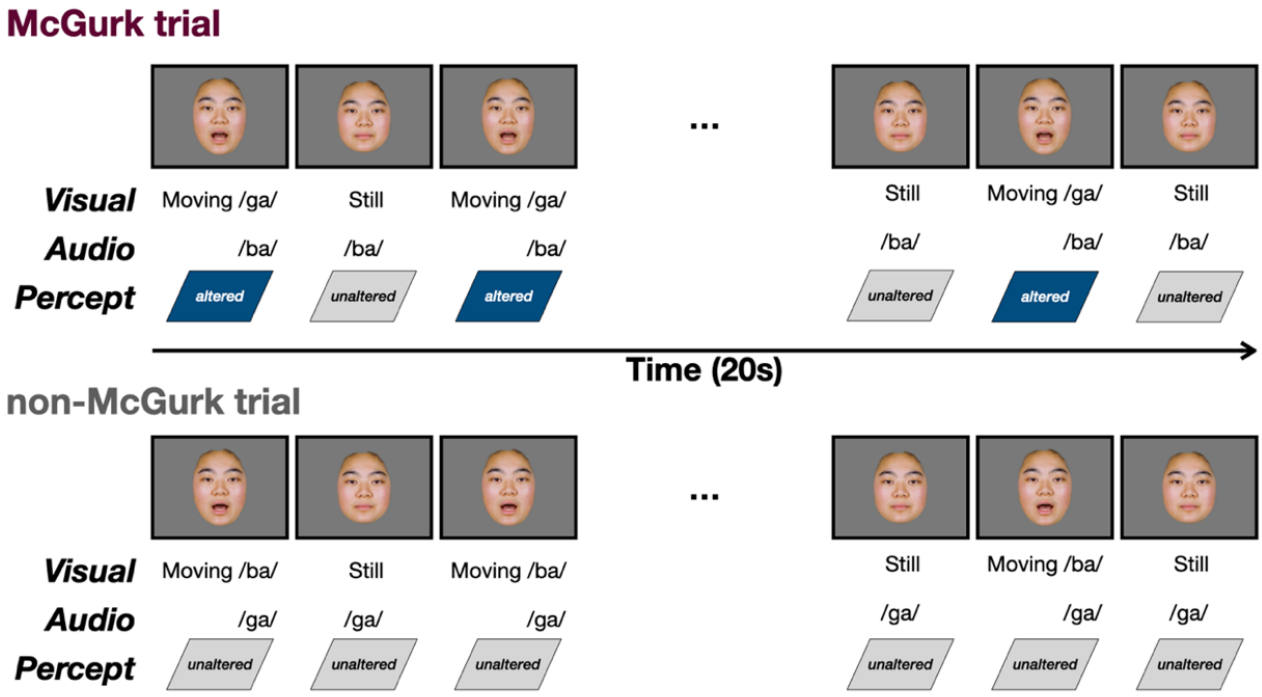


Figure 5. *Schematic presentation of the experimental procedures in the McGurk and non-McGurk trials (Experiment 2; other-race condition).* One among the eight East-Asian stimulus

[4] These eight non-White primary caregivers self-identified as follows: Arabic (n = 1), biracial (n = 1; Black and White), East-Asian (n = 1), Filipino (n = 3), and Indian (n = 3). The rationale for exclusion will be discussed in more details in the *General Discussion* section.

faces and two among the four audiovisual syllable pairs (i.e., visual /ga/-auditory /ba/ and visual /ba/-auditory /ga/) are presented here for demonstration purposes.

Results and Discussion.

Infants showed evidence of audiovisual speech integration.

The participants completed an average of 14.4 trials. As in Experiment 1, to examine whether White Canadian infants exhibited audiovisual speech integration when viewing the other-race (East-Asian) faces, we conducted a paired-sample *t*-test comparing the participants' average total looking times in McGurk and non-McGurk trials. As shown in Figure 6, the results revealed a significant looking preference ($t(20) = -3.29$, $p = .004$, Cohen's $d = 0.72$), suggesting that infants' auditory perceptual outcomes differed notably between the two types of trials, which constitutes evidence of audiovisual integration. However, in contrast to Experiment 1, infants looked significantly longer at the non-McGurk trials ($M = 11.78$ s, $SD = 3.19$) than at the McGurk trials ($M = 10.79$ s, $SD = 3.11$). This reversed looking preference indicates that infants' integration processes differ when viewing other-race versus own-race faces.

Prior to performing the paired-sample *t*-test, we verified the assumption of normality of each infant's McGurk preference scores (looking times to the McGurk trials – to non-McGurk trials) using the Shapiro–Wilk test, which indicated no deviation from normality ($W = 0.97$, $p = .784$). Visual inspections of the histogram and Q-Q plot supported the normality assumption.

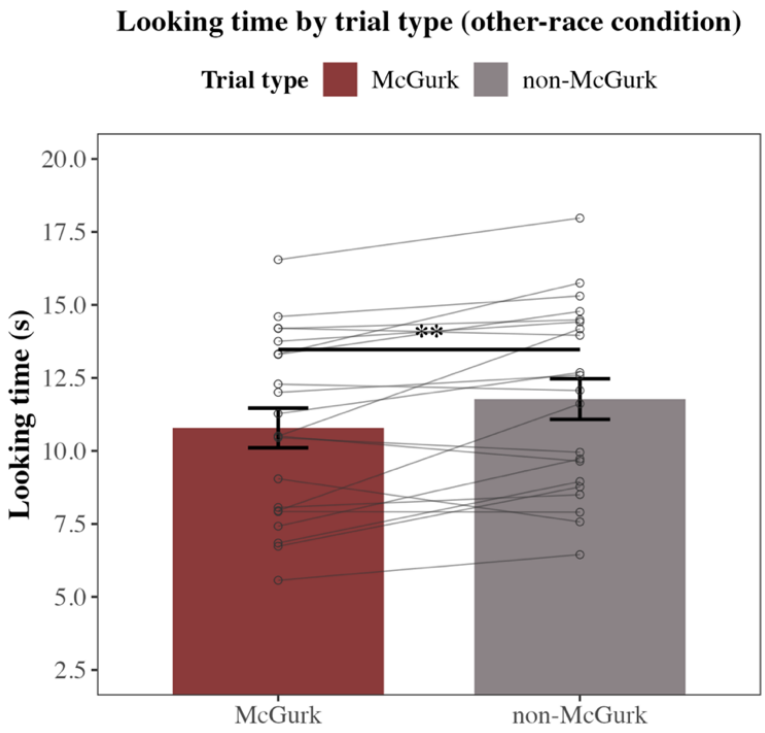


Figure 6. *Mean looking times for the McGurk and non-McGurk trials in Experiment 2.* The asterisks represent the statistically significant difference in looking time between two types of trials (** $p < .01$). Error bars represent ±1 standard error from the mean. Each dot represents an individual participant's looking time in each trial type, and the lines connecting paired dots reflect within-participant comparisons.

No age-related changes in audiovisual speech integration across infancy.

To assess whether there were any developmental shifts underlying infants' behavioural manifestation of the McGurk effect, we examined the correlation between infants' age measured in days and the preference scores for the McGurk trials, which was calculated in the same way as in Experiment 1. Pearson's correlation was not significant ($r(19) = –.14$, $p = .555$; Figure 7), indicating that infants consistently looked longer at the non-McGurk trials throughout the latter half of the first year.

Prior to the correlation analysis, we confirmed that all assumptions for Pearson's *r* were met. The relationship between age and the difference scores indicating McGurk trial preference was linear based on visual inspection of the scatterplot. The residuals from the linear model (*diff ~ ageDays*) were normally distributed ($W = 0.97$, $p = .676$), and there was no evidence of heteroscedasticity. Cook's distance analysis identified no influential observations.

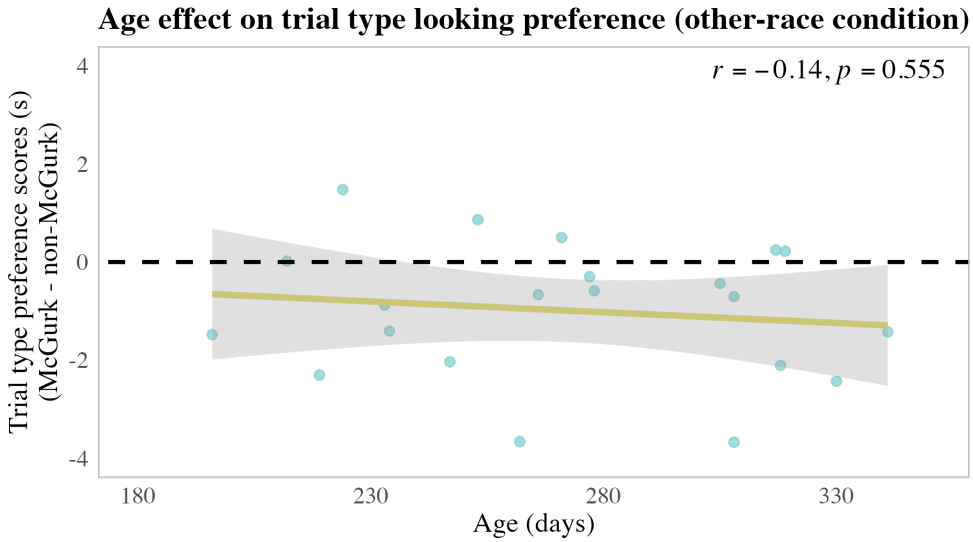


Figure 7. *Age-related change in trial looking preference in other-race face condition (Experiment 2).* Each dot represents data from an individual participant. The horizontal dashed line indicates equal looking to McGurk and non-McGurk trials. Dots above and below the dashed line are individuals who looked longer to the McGurk and non-McGurk trials, respectively. The shaded area around the regression line reflects the confidence interval of the regression model.

These findings suggest that, in the other-race face condition, infants might have perceived the McGurk trials as containing irregular and unpredictable alternations of two sounds (see Figure 2c)—a pattern from which they disengaged, instead showing increased looking to the repetitive and thus more predictable non-McGurk trials (see Figure 2b). This suggested that

when viewing other-race faces, infants could occasionally—but not reliably—integrate auditory and visual speech syllables when they were fusible, appearing less able to engage in integration compared to when seeing own-race faces. This indicated a potential disruption of integration in the presence of racially unfamiliar faces. Furthermore, the absence of any age-related changes suggests that this preference for non-McGurk trials remained stable across the tested age range, implying that other-race faces may begin to interfere with integration as early as 6 months of age—coinciding with the onset of perceptual narrowing in unimodal face processing (e.g., Kelly et al., 2007, 2009). Taken together, these findings support our second hypothesis that audiovisual speech integration is weakened by other-race faces in the latter half of the first year, revealing an other-race effect (ORE) in bimodal speech perception. Notably, and in contrast to prior findings reporting a complete lack of audiovisual integration in the presence of other-race faces (Ujiie et al., 2020, 2021), our data still revealed evidence of integration—albeit in a diminished form.

Infants' looking behaviours in Experiments 1 and 2

Trial Looking Preference Analysis

To more directly investigate the discrepancy in infants' audiovisual speech integration when seeing other-race faces versus own-race faces, we combined data from all infants in Experiments 1 and 2. A Welch's two-sample *t*-test was performed to compare infants' McGurk trial preference scores—the difference in looking time between McGurk and non-McGurk trials, which served as a proxy of their audiovisual speech integration capacity in our study—across the own-race and other-race face conditions.

The analysis revealed a significant difference in infants' trial looking preference between the two face-race conditions ($t(40.29) = -4.08$, $p < .001$). Specifically, infants in the own-race condition showed a positive McGurk preference ($M = 0.53$), indicating greater looking toward

the McGurk trials relative to non-McGurk trials, which indexed a developed integration capacity. In contrast, infants in the other-race condition exhibited a negative preference score ($M = –0.99$), reflecting longer looking to non-McGurk than McGurk trials—suggesting an underdeveloped integration. These findings provide more direct evidence that other-race faces attenuate infants' audiovisual speech integration, substantiating an other-race effect (ORE) in multisensory speech perception.

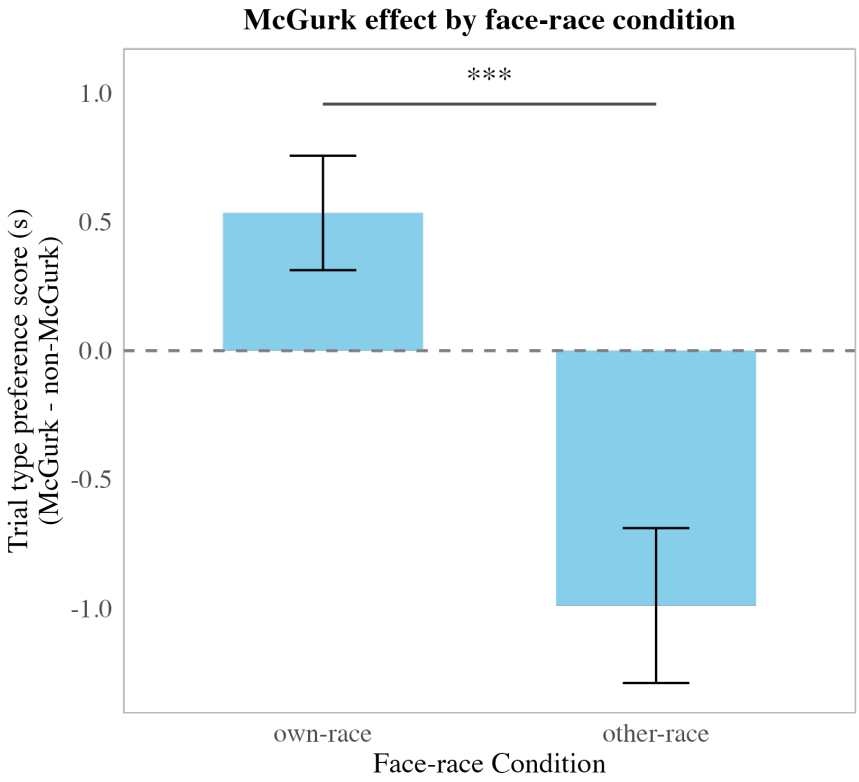


Figure 8. *Mean McGurk trial preference scores in own-race (Experiment 1) and other-race (Experiment 2) conditions.* The asterisks signify the statistically significant difference in looking time between two trial types (*** $p < .001$). The horizontal dashed baseline represents an equal looking time to the McGurk and non-McGurk trials. Bars above and below the baseline indicate longer looking to the McGurk and non-McGurk trials. Error bars represent ±1 standard error from the mean.

Given the observed interference in audiovisual speech integration when infants viewed other-race faces (i.e., ORE), we next examined their fine-grained face-scanning patterns to explore potential differences in visual attention allocation between own-race and other-race conditions—an explanation previously proposed to account for the ORE in bimodal speech integration (Ujiie et al., 2021). This analysis was based on the hypothesis that such disruption in integration might stem from differential attentional mechanisms infants employ when viewing own-race versus other-race faces. Specifically, we began with analyzing infants' proportional looking to the mouth, as prior research indicates that increased mouth-looking plays a crucial role in enhancing audiovisual speech integration (e.g., Gurler et al., 2015; Kushnerenko et al., 2013; Stacey et al., 2020). Accordingly, we hypothesized that infants would exhibit greater proportional mouth-looking in the own-race condition than in the other-race condition.

Areas-of-Interest (AOIs) Analysis

Since our primary aim was to examine how face-race familiarity may influence infants' overall allocation of visual attention to different areas-of-interest (AOIs), we collapsed data across both trial types (i.e., McGurk and non-McGurk) within each face-race condition for the subsequent AOI analyses. Although linear mixed-effects models indicated that trial type systematically influenced PTLT-mouth, our focus was not on trial-level differences, but rather on broader, condition-level patterns of face-scanning that may help explain the observed ORE in audiovisual speech integration (see Figure 8). Additionally, the influence of trial type appeared to be global across both conditions, as indicated by the absence of interaction effects, improving comparability and further justifying this analytical choice of collapsing. Moreover, trial type was evenly distributed across blocks and conditions (i.e., two McGurk and two non-McGurk trials

per block), reducing concerns about systematic bias. For detailed AOI analyses by trial type, see the *Supplementary Materials* (section: PTLT-Mouth by Trial Type and Face Condition).

PTLT-mouth did not significantly differ between face-race conditions.

The proportional looking time at the mouth is a well-established measure of infants' face-scanning strategies and has been linked to the development of speech and language capacities in both infancy (Lewkowicz & Hansen-Tift, 2012) and later childhood (Young et al., 2009). Therefore, it serves as the primary focus of our areas-of-interest (AOIs) analysis.

To ensure precision and consistency, we used a face template approach (Xiao & Lee, 2018) to define three AOIs: the mouth, the eyes, and the whole face, as shown in Figure 9a. This approach allowed for standardization across stimulus faces, minimizing variability introduced by differences in facial proportions or spatial arrangements.

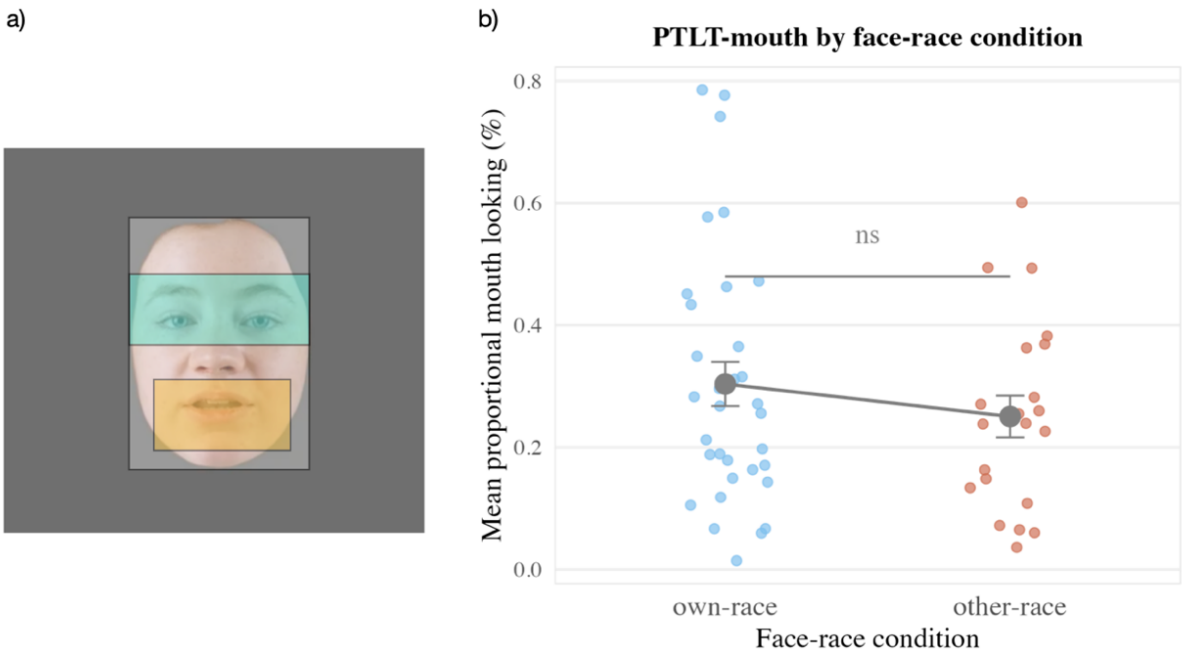


Figure 9. *Areas-of-Interest (AOI) definitions using a face template and infants' proportional-of-total-looking to the mouth (PTLT-mouth) in own-race and other-race conditions.* **a)** Definitions

of three areas of interest (AOIs) using the face template approach. The turquoise, yellow, and light grey regions represent the eyes, mouth, and the whole face area, respectively. All AOI analyses in this study followed these same definitions of AOIs. **b)** Proportional mouth-looking in both own-race (Experiment 1) and other-race (Experiment 2) face conditions. The blue and red dots represent infants' proportional mouth-looking in the own-race and other-race face condition, respectively.

To examine whether infants' attentional allocation to the mouth region differed by face-race condition, all infants' data from the two experimental conditions were combined. For each infant, we computed the proportion-of-total-looking (PTLT) to the mouth by dividing the total amount of looking directed at the mouth by the total amount of looking at any portion of the whole face as follows to index their visual attention to the mouth area:

$$PTLT_{mouth} = \text{Total looking time}_{mouth} / \text{Total looking time}_{whole\ face}$$

A Welch's two-sample *t*-test comparing the mean PTLT-mouth between the two face-race conditions revealed that infants did *not* look significantly more to the mouth when viewing own-race versus other-race speaking faces ($t(50.44) = 1.07$, $p = .289$; Cohen's $d = .28$; see Figure 9b), although the descriptive means suggested a trend in the predicted direction (own-race: $M = .30$; other-race: $M = .25$).

This result suggests that, at a group level, infants' overall allocation of visual attention to the mouth region is not systematically influenced by face-race. However, given the theoretical relevance of mouth-looking for audiovisual speech perception, we next explored whether the magnitude of the McGurk fusion effect was associated with PTLT-mouth. We hypothesized a significant positive correlation between McGurk strength and PTLT-mouth, such that stronger McGurk effects would be associated with longer proportional looking times to the mouth area.

Lack of associations between PTLT-mouth and McGurk integration magnitude.

In this section, we examined whether infants' proportional looking to the mouth (PTLT-mouth) was associated with the strength of their audiovisual speech integration, operationalized by their McGurk trial preference scores (i.e., the difference between average total looking times to the McGurk and non-McGurk trials). Larger and more positive scores indicate stronger integration.

A Pearson correlation using the combined dataset from Experiments 1 and 2 revealed no significant association between McGurk preference scores and PTLT-mouth ($r(52) = -.02$, $p = .893$), implying that stronger McGurk fusion does not appear to systematically relate to greater PTLT-mouth in the present dataset. Therefore, PTLT-mouth alone is unlikely to convincingly account for the observed discrepancy in audiovisual integration between the two conditions (i.e., the ORE). When analyzed separately by face-race condition, a marginally significant negative correlation emerged in the own-race condition ($r(31) = -.30$, $p = .086$; Figure 10, left panel), suggesting evidence that infants who showed stronger integration tended to look proportionally *less* at the mouth region of own-race speakers. This finding further indicates that longer overall PTLT-mouth may not correspond to more robust fusion effects. In contrast, in the other-race condition, the correlation between PTLT-mouth and the McGurk preference scores was positive yet far from statistical significance ($r(19) = .30$, $p = .185$; Figure 10, right panel).

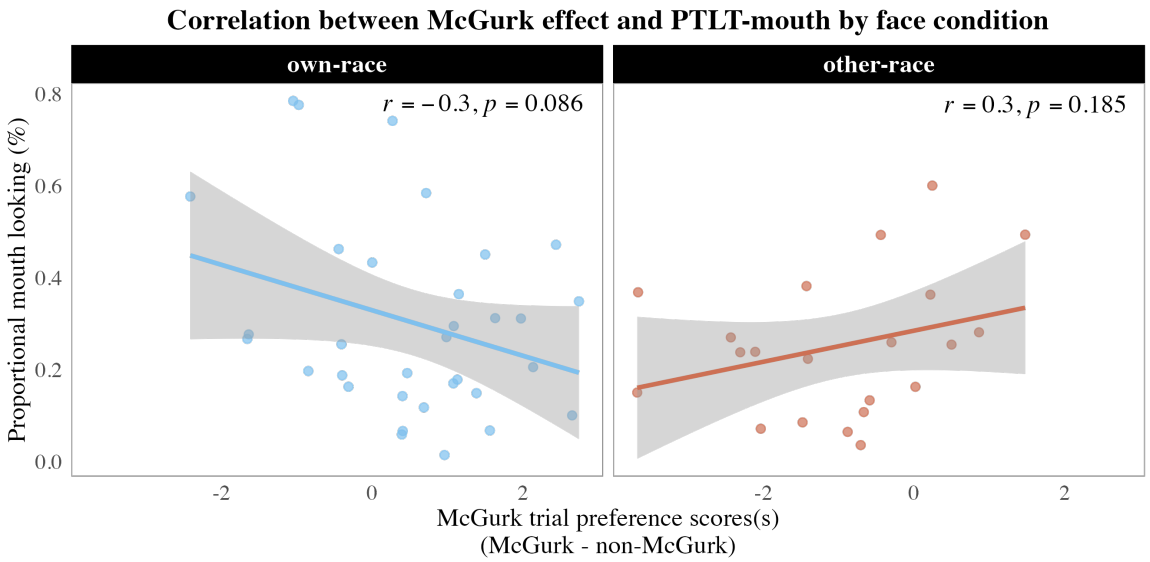


Figure 10. *Correlations between the magnitude of McGurk effect (McGurk preference scores) and PTLT-mouth by own-race and other-race conditions.* **a)** Left panel: a marginally significant negative correlation between McGurk trial preference scores and PTLT-mouth looking in the own-race face condition, suggesting that infants who showed a stronger McGurk effect looked proportionally less to the mouth region. Each blue dot represents an individual infant's PTLT-mouth looking and the corresponding magnitude of the McGurk effect. **b)** Right panel: a non-significant positive correlation between McGurk trial preference scores and PTLT-mouth looking in the other-race condition. Each red dot indicates an infant's PTLT-mouth looking and corresponding strength of the McGurk fusion.

This pattern of findings may appear counterintuitive given that greater attention to the mouth has been associated with enhanced audiovisual speech perception (Gurler et al., 2015; Kushnerenko et al., 2013). One plausible interpretation is that, when viewing own-race faces, infants who showed stronger integration may have required relatively less visual support from the articulating mouth; conversely, those who looked more at the mouth may have relied more heavily on visual speech cues to compensate for a weaker internal processing of bimodally

incongruent speech information and to aid the integration process. Despite that this observed correlation was marginal did not reach the conventional significance threshold, the trend suggests an interesting possibility of the relation between infants' mouth-looking and their capacity to integrate visual speech cues to their auditory perception, inviting further investigations. The lack of a clear association between the strength of McGurk fusion and PTLT-mouth in other-race condition may reflect a less consistent use of visual strategies when processing unfamiliar faces.

Together, these findings yielded two implications. First, longer PTLT-mouth does not equate to a stronger audiovisual speech integration, suggesting that examining PTLT-mouth in isolation may not adequately account for the integration differences observed across face-race conditions. Second, face-race familiarity *may* influence the relationship between visual attention to the source of auditory speech information and the integration of bimodal speech cues in infancy. More broadly, the differing patterns of correlation between integration strength and PTLT-mouth may underscore how face-race familiarity shapes the functional role of visual attention in supporting the integration process. However, we emphasize that these interpretations should be treated with caution and considered speculative, given the absence of statistically significant correlations between PTLT-mouth and integration strength.

So far, we have not identified a convincing explanatory account within the same dataset for the observed ORE based on PTLT-mouth alone. To further investigate why other-race faces interfere with audiovisual speech integration, we next examined how infants distributed visual attention to both the mouth and the eyes regions in the real-time—zooming in dynamic changes across the four experimental blocks—to further elucidate mechanisms that may underlie the ORE in bimodal speech integration. This analysis was theoretically motivated by prior research (e.g.,

Liu et al., 2013; Wheeler et al., 2011) demonstrating that infants engage in differential face-scanning strategies when viewing own- versus other-race faces.

Changes in visual attention across blocks reveal race-dependent strategy shifts.

Building on our previous finding that the overall PTLT-mouth did not differ significantly across own- and other-race conditions, we next asked whether infants' visual attention patterns changed dynamically over the course of the experimental session, and whether these changes differed by face-race conditions. This analysis aimed to capture whether infants flexibly adjusted their face-scanning strategies in real time—an adaptation that could support audiovisual speech integration. In addition to the mouth as a crucial AOI for speech perception, we included PTLT eyes in this analysis to provide a fuller picture of infants' visual attention allocation, as the eyes region is central to both face recognition and the interpretation of social signals (e.g., Farroni et al., 2005; Schyns et al., 2002; Hadjikhani et al., 2008) and is particularly relevant here given that racial group membership functions as a salient social cue for infants (e.g., Yuan et al., 2019).

To maintain consistency with the primary McGurk effect analyses, we included data from all four experimental blocks, each comprising four trials (two McGurk and two non-McGurk). Including the full experimental span allowed us to capture the complete trajectory of visual engagement and assess whether attentional shifts emerge incrementally with exposure. Worth noting, some infants contributed only partially to later blocks (e.g., completing 1–3 trials in Block 3 or Block 4; see Table 2 for details of participant attrition). To accommodate this natural attrition while maximizing usable data, we employed linear mixed-effects models, which effectively account for partial trial contributions, unequal group sizes, and repeated measures of blocks by modeling participant ID as a random intercept.

We fit separate models for PTLT-mouth and PTLT-eye, including face-race condition (own vs. other), block number (treated as a categorical variable), and their interaction as fixed effects. This modeling framework allowed us to test whether visual attention to different facial regions evolved over time, and whether this tuning was modulated by the racial familiarity of speakers. Such analysis provides us a lens through which to probe the mechanisms underlying the other-race effect (ORE) observed in audiovisual speech integration.

Table 2. The attrition patterns of participants on the block and trial levels across face-race conditions.

Face-race condition	Block No.	Total included (≥1 trial) Infant No.	Completed all 4 trials Infant No.	Partial trials (2-3 trials) Infant No.
Own-race	1	33	33	0
	2	33	33	0
	3	33	30	3 (2 with 2 trials, 1 with 3)
	4	27	23	4 (2 with 2 trials, 2 with 3)
Other-race	1	21	21	0
	2	21	21	0
	3	20	18	2 (both with 2 trials)
	4	17	13	4 (2 with 2 trials, 2 with 3)

PTLT-mouth across blocks in both race conditions.

The linear mixed-effects model for PTLT-mouth revealed a significant main effect of block number ($F(3, 145.94) = 12.80, p < .001$) with a large effect size ($\eta^2_p = .21$), indicating that infants' attention to the mouth decreased over time. Neither the main effect of face-race

condition ($F(1, 52.49) = 0.96$, $p = .33$; $\eta^2_p = .02$) nor the interaction between face-race and block number ($F(3, 145.94) = 0.42$, $p = .74$; $\eta^2_p < .01$) reached significance, suggesting a global decline in mouth-looking across both race conditions.

Follow-up pairwise comparisons further clarified this pattern within each face-race condition. In the own-race condition, PTLT-mouth showed a numerically decreasing—though not significant—trend across adjacent blocks (Block 1 vs. 2: estimate = 0.052, $p = .196$; Block 2 vs. 3: estimate = 0.060, $p = .104$; Block 3 vs. 4, estimate = 0.029, $p = .729$), with significant drops observed between more distant timepoints (Block 1 vs. 3: estimate = 0.112, $SE = 0.026$, $p < .001$; Block 2 vs. 4: estimate = 0.089, $SE = 0.028$, $p = .010$). By the end of the session, infants' PTLT-mouth was significantly reduced compared to the beginning (Block 1 vs. 4: estimate = 0.141, $SE = 0.028$, $p < .001$). In the other-race condition, PTLT-mouth also exhibited a pattern of numeric decrease across adjacent blocks (Block 1 vs. 2: estimate = 0.066, $p = .186$; Block 2 vs. 3: estimate = 0.018, $p = .952$; Block 3 vs. 4, estimate = 0.035, $p = .764$), with a marginally significant decline from Block 1 to 3 (estimate = 0.084, $SE = 0.033$, $p = .062$) and a significant drop from Block 1 to 4 (estimate = 0.118, $SE = 0.035$, $p = .006$). This overall pattern suggests that while infants in both face conditions reduced attention to the mouth over time, pronounced declines occurred earlier and was more consistently detectable in the own-race condition. Figure 11 visualizes the comparisons illustrated above.

Between-condition comparisons at each block revealed no significant differences in PTLT-mouth between infants viewing own- versus other-race faces (Block 1: estimate = 0.062, $p = .300$; Block 2: estimate = 0.076, $p = .204$; Block 3: estimate = 0.034, $p = .575$; Block 4: estimate = 0.039, $p = .526$), although infants looked numerically more to the mouth in own-race

condition. These results indicated that, at each discrete time point, infants allocated comparable levels of attention to the mouth region regardless of the race of the faces.

PTLT-eye across blocks in both race conditions.

The linear mixed-effects model for PTLT-eye revealed no main effects of block ($F(3, 145.84) = 0.72$, $p = .543$; $\eta^2_p = .01$) or face-race condition ($F(1, 52.35) = 0.15$, $p = .700$; $\eta^2_p < .01$), indicating that infants' overall attention to the eye region did not differ across blocks or by the races of stimulus faces, when considered independently. However, a significant interaction between block number and face-race condition emerged ($F(3, 145.84) = 2.86$, $p = .039$; $\eta^2_p = .06$), suggesting that the temporal pattern of eye-looking may vary by face-race condition.

Follow-up pairwise comparisons clarified this interaction effect. In both own-race and other-race conditions, no significant differences were observed across any pair of blocks (all ps > .214), suggesting that infants' attention to the eyes remained stable across the study session. In the own-race condition, PTLT-eye showed a numerically increasing—though not statistically significant—trend across adjacent blocks (Block 1 vs. 2: estimate = –0.016, $p = .945$; Block 2 vs. 3: estimate = –0.038, $p = .574$; Block 3 vs. 4, estimate = –0.007, $p = .996$), with the largest contrast occurring between Block 1 and Block 4 (estimate = –0.061, $p = .213$). In the other-race condition, none of the block-wise comparisons of PTLT-eye reached statistical significance (all ps > .172), suggesting a similarly stable pattern across the course of experiment. However, unlike in the own-race condition, infants' PTLT-eye did not follow a consistent pattern, with an increase from Block 1 to 2 (estimate = –0.052, $p = .495$), a decline from Block 2 to 3(estimate = 0.077, $p = .172$), and another minor increase from Block 3 to 4 (estimate = –0.005, $p = .999$). The largest contrast between Block 2 and Block 4 (estimate = 0.072, $p = .264$). While none of these contrasts were statistically significant, this pattern may suggest that infants maintain a relatively stable

level of eye-looking across the experiment session, particularly in own-race condition, whereas their attention to the eyes fluctuates slightly when viewing unfamiliar-race faces (see Figure 11).

Between-condition comparisons at each block yielded no significant differences in PTLT-eye (Block 1: estimate = –0.053, p = .406; Block 2: estimate = –0.089, p = .169; Block 3: estimate = 0.026, p = .169; Block 4: estimate = 0.028, p = .676), suggesting that infants allocated comparable attention to the eye region regardless of face-race familiarity at any given block. Nonetheless, a subtle numerical pattern emerged: infants appeared to initially direct more attention to the eyes of other-race speakers in the earlier blocks (Blocks 1 and 2), yet this pattern reversed in later blocks (Blocks 3 and 4), with slightly more eye-looking to own-race speakers.

Though speculative and not supported by significant effects, these observations warrant further investigations into the temporal dynamics of infants' visual attention to faces of familiar and unfamiliar races. In the following section, we turn to the within-block comparisons of PTLT-mouth and PTLT-eye to assess how infants distributed visual attentional resources between the mouth and the eyes when processing own-versus other-race speaking faces.

Comparisons between PTLT-mouth and PTLT-eyes within blocks.

To probe the visual attention mechanisms that might underlie the other-race effect (ORE) observed in audiovisual speech integration, we conducted within-block comparisons of infants' proportional looking time (PTLT) to the mouth and eyes. This analysis aimed to determine whether infants favoured one region over the other at different timepoints and whether this preference varied depending on face-race familiarity—factors that could influence the quality of audiovisual speech processing. We approached this analysis by directly comparing PTLT-mouth and PTLT-eye using paired-sample t-tests within each block and face-race condition (Figure 11).

In the own-race condition, no significant differences emerged in Block 1 ($t(32) = –0.16$, $p = .877$, Cohen's $d = –0.03$) and Block 2 ($t(32) = –1.05$, $p = .303$, Cohen's $d = –0.18$). However, infants looked significantly more to the eyes than to the mouth in Block 3 ($t(32) = –2.29$, $p = .029$, Cohen's $d = –0.40$) and marginally so in Block 4 ($t(26) = –1.95$, $p = .062$, Cohen's $d = –0.38$). This later shift may reflect a gradual reallocation of visual attention toward the eye region as infants became more familiar with the audiovisual speech stimuli, adapted their face-scanning strategies over time, or experienced increasing fatigue.

In the other-race condition, no significant difference was found in Block 1 ($t(20) = –1.38$, $p = .184$, Cohen's $d = –0.30$). However, starting in Block 2, infants attended significantly more to the eyes than the mouth ($t(20) = –3.92$, $p < .001$, Cohen's $d = –0.85$), with this pattern persisting in Block 3 ($t(19) = –2.30$, $p = .033$, Cohen's $d = –0.51$) and Block 4 ($t(16) = –2.12$, $p = .050$, Cohen's $d = –0.51$). These findings suggest that when faced with unfamiliar-race faces, infants may more readily default to scanning the eye region. While speculative, this early-emerging and stable preference for the eyes could reflect heightened social monitoring or reduced reliance on the articulatory speech cues when perceiving racially unfamiliar faces—both of which may limited opportunities for successful audiovisual speech integration.

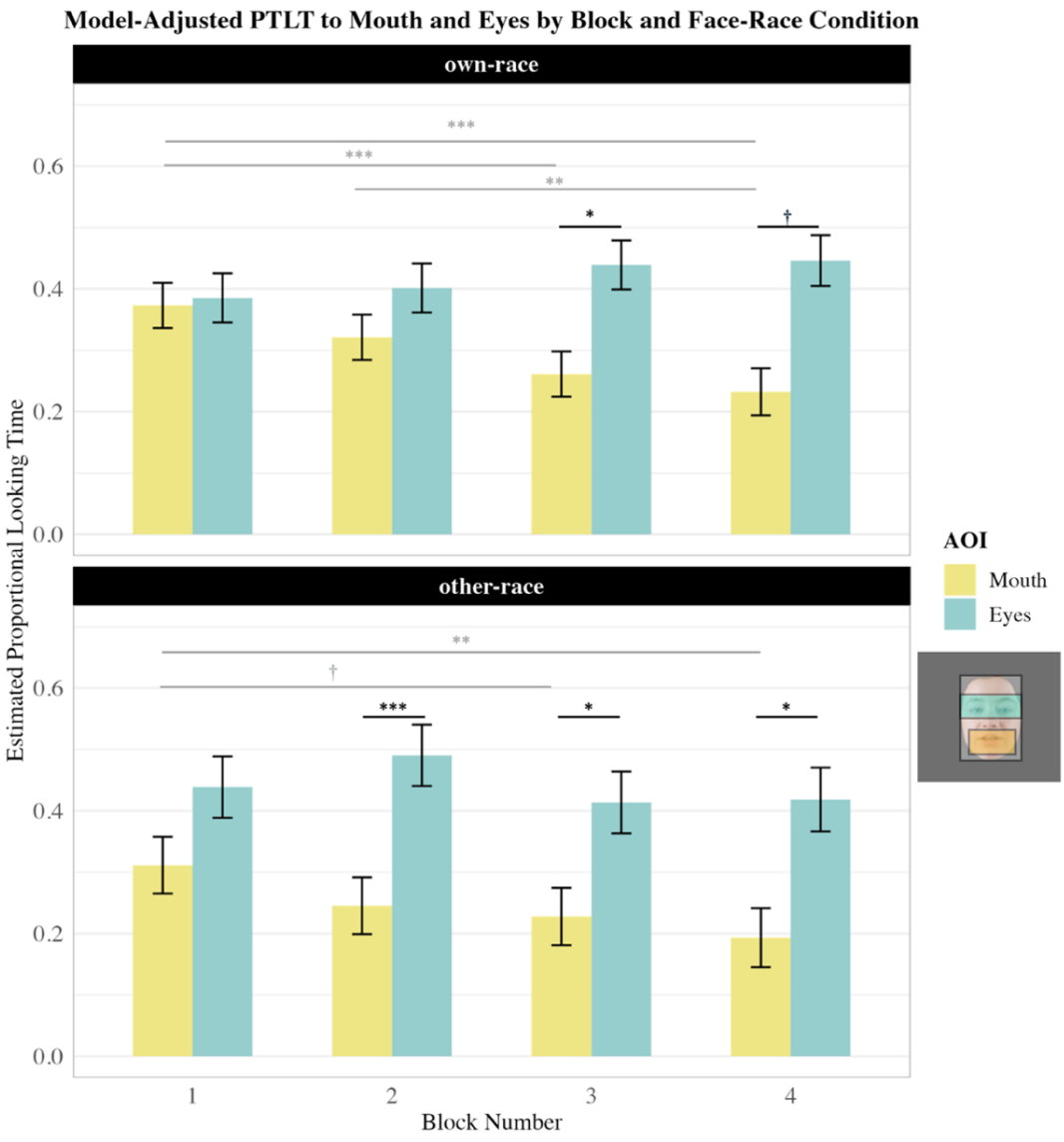


Figure 11. *Infants' PTLT-mouth and PTLT-eyes in all experimental blocks in the own-race and other-race face condition.* **The upper panel:** infants' PTLT to the mouth (yellow bar) and the eyes (turquoise) in the own-race condition (Experiment 1). The asterisks represent statistically significant differences within or across blocks (* $p < .05$; ** $p < .01$; *** $p < .001$). The error bars represent ±1 standard error from the mean. **The lower panel:** infants' PTLT to the mouth and eyes in the other-race condition (Experiment 2). The dagger "†" indicates that the differences between PTLT-mouth across blocks 1 and 3 are marginally significant ($.05 < p < 1.0$). The asterisks signify statistically significant differences within or across blocks.

These within- and between-block-level analyses might reveal a nuanced picture of how race familiarity shapes infants' visual attention in ways that impair or support audiovisual speech integration. Specifically, although face-race familiarity did not systematically impact infants' average PTLT-mouth (Figure 9b), it may modulate infants' dynamic allocation of attention to specific facial regions—namely, the mouth, an AOI essential with articulatory speech cues, and the eyes, an AOI central to face recognition and social-communicative cues—over the course of the session. These differing patterns of real-time attentional change across face-race conditions, in turn, may contribute to the observed ORE in audiovisual speech integration (Figure 8).

We discuss tentative explanations for the observed ORE based on our AOI analysis of infants' dynamic allocation of visual attention. When comparing PTLT-mouth and PTLT-eye within blocks over time, infants' earlier and more consistent preference for the eyes over the mouth when viewing other-race faces may reflect a reduced prioritization of articulatory speech cues, potentially undermining robust integration of auditory and visual speech cues. This interpretation aligns with prior work suggesting that selective attention to the mouth over the eyes supports infants to extract and process the audiovisual speech information (e.g., Lewkowicz & Hansen-Tift, 2012; see Pons et al., 2015, Birulés et al., 2019, for evidence from bilingual infants). This significantly lower PTLT-mouth compared to PTLT-eye may arise for two reasons. First, infants may seek ostensive social-pragmatic cues (e.g., communicative intent) when exposed to individuals from an unfamiliar race, which they may have perceived as a salient indicator of different community membership (Uttley et al., 2013; Weatherhead & Werker, 2022)—reflecting a *voluntary* and notable preference for the eyes over the mouth. Second, infants may have found the mouth region of other-race speakers less perceptually informative for extracting visual speech cues (see Experiment 2 in Ujiie & Takahashi, 2022, for evidence from

adults), leading them to prioritize the eyes either as a more consistent source of social information or as a default target when audiovisual binding fails, reflecting a *compulsory* preference for the eyes over the mouth.

In comparison, infants viewing own-race faces initially showed no strong bias toward the eyes over the mouth, only directing significantly more attention to the eyes in the last two blocks. While speculative, the delayed onset of such notable attentional shift may have afforded infants more chance to extract and incorporate articulatory information from the mouth into their auditory perception, thereby facilitating more robust audiovisual integration. Another possibility is that when viewing own-race faces enacting the McGurk stimuli, infants prioritized the eyes significantly only after acquiring sufficient articulatory inputs from the mouth—as indicated by the significant drop in PTLT-mouth from Block 1 to Block 3—which may have contributed to the emergence of a statistically reliable eye preference by Block 3. This pattern may reflect a habituation-like response, wherein attention shifted increasingly toward the eyes once the mouth became more predictable or less informative—not as a categorical shift, but as an amplification of an existing bias. This pattern again suggests that infants' face-scanning strategies may be dynamically modulated by stimulus familiarity and information demands.

To sum up, using the McGurk effect as a proxy measure and infants' looking preferences for the McGurk and non-McGurk trials as the dependent variable, the first two experiments observed evidence of early audiovisual integration in White Canadian infants, though arguably with different strengths depending on the races of faces presenting the speech stimuli. From six months onward, infants demonstrate a *robust and stable* integration when viewing McGurk audiovisual stimuli enacted on own-race faces (Experiment 1), as indicated by their significantly longer looking times to the McGurk than to non-McGurk trials across the latter half of the first

year. This pattern suggested that infants might have perceived a regular alternation between the fused auditory percept (e.g., /da/ or /ta/) and the physical auditory syllable (e.g., /ba/ or /pa/) during McGurk trials, making the auditory stream more structured, predictable, and perceptually engaging. In contrast, when the same McGurk and non-McGurk stimuli were enacted on other-race (East-Asian) faces (Experiment 2), infants looked consistently longer at the non-McGurk trials across age. This preference indicated that infants might have perceived sporadic and inconsistent sound sequences that randomly switched between the fused percepts and physical auditory syllables, which led them to prefer the more predictable repetition in the non-McGurk trials. This implied that infants showed an *under-developed* audiovisual speech integration when viewing other-race faces from 6 to 12 months of age, suggesting an other-race effect (ORE).

Critically, these interpretations hinge on the inferred patterns of auditory perceptual outcomes in the McGurk trials, which we derived by contrasting them with the perceptually stable and repetitive non-McGurk trials, allowing for theoretical comparisons across both experiments. Specifically, we proposed two hypothetical perceptual outcomes: one in which infants perceived a regularly alternating auditory sequence (Experiment 1), and another in which they perceived a sporadically alternating sequence (Experiment 2), both compared against the repetitive auditory stream in the non-McGurk trials. This inference was grounded in prior research on infants' preferences for different levels of perceptual complexity (i.e., the Goldilocks effect; Kidd et al., 2012, 2014), which suggests that infants gravitate towards learnable patterns that strike a balance between unpredictability, repetition, and alternation.

However, since the perceptual outcomes in the McGurk trials could *not* be directly assessed, these hypotheses remained speculative. To better test these hypotheses and support our interpretations from Experiments 1 and 2, Experiment 3 aims to examine whether infants prefer

regularly alternating over repetitive sound patterns. In this experiment, infants were presented with actual alternating and non-alternating auditory sequences, with the articulatory mouth movements always being congruent with the dubbed auditory syllables. This approach removed the audiovisual incongruency characteristic of the McGurk effect, allowing us to assess—using infants looking preference as index—their preference for alternating and non-alternating audiovisual speech cues based purely on the patterned versus repetitive physical auditory sounds, rather than on potential interactions between conflicting visual and auditory information. Furthermore, Experiment 3 exposed participants to either own-race (White) or other-race (East Asian) faces in a counterbalanced design, to examine whether race continues to modulate infants' audiovisual perception when the speech signals from the two modalities are always matched.

Experiment 3

Participants.

Twenty-three full-term Canadian infants (12 females) with normal vision and hearing participated in the current experiment after caregivers provided informed consent. The participants were from 192 to 379 days old (M = 9.84 months, SD = 1.67 months) and were recruited from the Southern Ontario area in Canada. Among the twenty-three participants, four had primary caregivers who were East-Asian, one had a primary caregiver who was Indian, one who was Black, and the remaining fifteen had caregivers who were White.

Six additional infants participated but were excluded from data analysis because of failure to complete the experimental procedure due to fussiness (n = 3), sleepiness (n = 2), and calibration failure (n = 1). No participants were removed based on their primary caregivers' race, as we hypothesized that face-race would not impact infants' auditory processing when the audiovisual speech signals are consistently congruent, which was later confirmed (see upcoming

Result and Discussion). To maintain consistency of the study design, infants with East-Asian primary caregivers were categorized based on the face-race they viewed: those who viewed East-Asian faces were assigned to the own-race condition (n = 1), and those who viewed White faces were assigned to the other-race condition (n = 3). The infants with Indian and Black primary caregivers both viewed a White face and was therefore categorized in the other-race condition (n = 2). The categorization of face-race condition for White Canadian infants was consistent with Experiments 1 and 2: infants who viewed White faces were labeled as own-race (n = 7), and those who viewed East-Asian faces as other-race (n = 10). This categorization criterion resulted in eight participants in the own-race condition and fifteen participants in the other-race condition.

All experimental protocols, including the procedures for obtaining the informed consent, were reviewed and approved by the Research Ethnics Board of McMaster University. The families received a T-shirt or a tote bag of their own choice in exchange for their participation.

Stimuli.

The raw materials for the auditory and visual stimuli used in Experiment 3 were identical to those used in Experiments 1 and 2. Specifically, all auditory (i.e., auditory syllables) and visual (i.e., visual syllables) components were created from recordings of the same prototype young female articulating the four syllables: /ba/, /pa/, /ga/, and /ka/. Each syllable was articulated for approximately 800 ms, and the recordings were trimmed into 1 second clips.

For the ***auditory syllables***, Adobe Audition was used to eliminate background noises and matched the overall loudness across the four audio syllables. The ***visual syllables*** were created by transferring the prototype actress's facial movements onto 16 new female faces (eight White and eight East-Asian) using GANs, producing photo-realistic faces that made identical

articulatory movements as the prototype actress. All stimulus faces were then standardized by removing hair, equalizing the face sizes, and placing them against a light gray background.

However, the procedural composition and presentation of these raw experimental stimuli differed from the previous two experiments in two ways, as will be detailed in the following *Procedure* section.

Procedure.

Similar to the procedural paradigm employed in Experiments 1 and 2, infant participants watched videos of faces articulating visual syllables while simultaneously hearing the auditory syllables across a maximum of *eight* (instead of sixteen) trials. Each trial still consisted of twenty iterations presented at the pace of one per second, lasting a total of twenty seconds.

Despite this similarity, two major changes were made. *First*, instead of presenting conflicting audio and visual syllables while the faces were moving (e.g., see the 1st, 3rd, 5th, 7th, 9th, 11th, 13th, 15th, 17th, and 19th iteration in Figures 1 and/or 5 regardless of trial type, i.e., McGurk vs. non-McGurk trials), the auditory and visual syllables in all trials of the current experiment were always congruent. This adjustment eliminated the audiovisual mismatch, as the goal was to examine infants' preference for alternating versus repetitive sound patterns, rather than their capacity to integrate conflicting audiovisual speech information. *Second*, instead of viewing a different face of the same race in each trial, each infant watched only one stimulus face articulating throughout the entire study session. This was intended to ensure that infants' visual recovery in a given trial was driven by the auditory pattern, rather than by novelty or visual salience of a new face.

The trials were divided into two types: Alternating and non-Alternating trials. In each ***Alternating trial***, infants watched one congruent audiovisual syllable (1 second) repeated 10 times. These ten video clips were interleaved with ten still-face images, created by holding the

first frame of the video recording of the visual syllable articulation. Each still-face image (i.e., the 2nd, 4th, 6th, 8th, 10th, 12th, 14th, 16th, 18th, and 20th iteration) lasted for 1 second and was dubbed with one of the three auditory syllables that differed from the one in the immediately preceding iteration (i.e., the 1st, 3rd, 5th, 7th, 9th, 11th, 13th, 15th, 17th, and 19th). As a result, infants viewed alternating sequences of a moving, articulating face and a still-face image, while hearing two sounds that alternated regularly throughout the trial. The auditory and visual syllables always matched when the face was moving. In each ***non-Alternating trial***, the overall structure remained the same, but the auditory syllable did not alternate across iterations. Infants continued to view alternating sequences of an articulating face and the still-face image of the same stimulus face, with this moving–still–moving–still visual stream paired with a single repeated auditory syllable. As in the *Alternating* trial, the auditory and visual syllables were always matched during the articulating face segments.

The perceptual auditory outcomes in both trial types should reflect the physical auditory syllables, as there was no audiovisual mismatch. Figure 12 schematizes the combinations of visual and auditory stimuli, along with the corresponding auditory perceptual outcomes for the Alternating and non-Alternating trials. The perceptual outcomes in the Alternating trials should align with those observed when the audiovisual integration is developed (Figure 2a), and those in the non-Alternating trials should align with those when the integration is absent (Figure 2b).

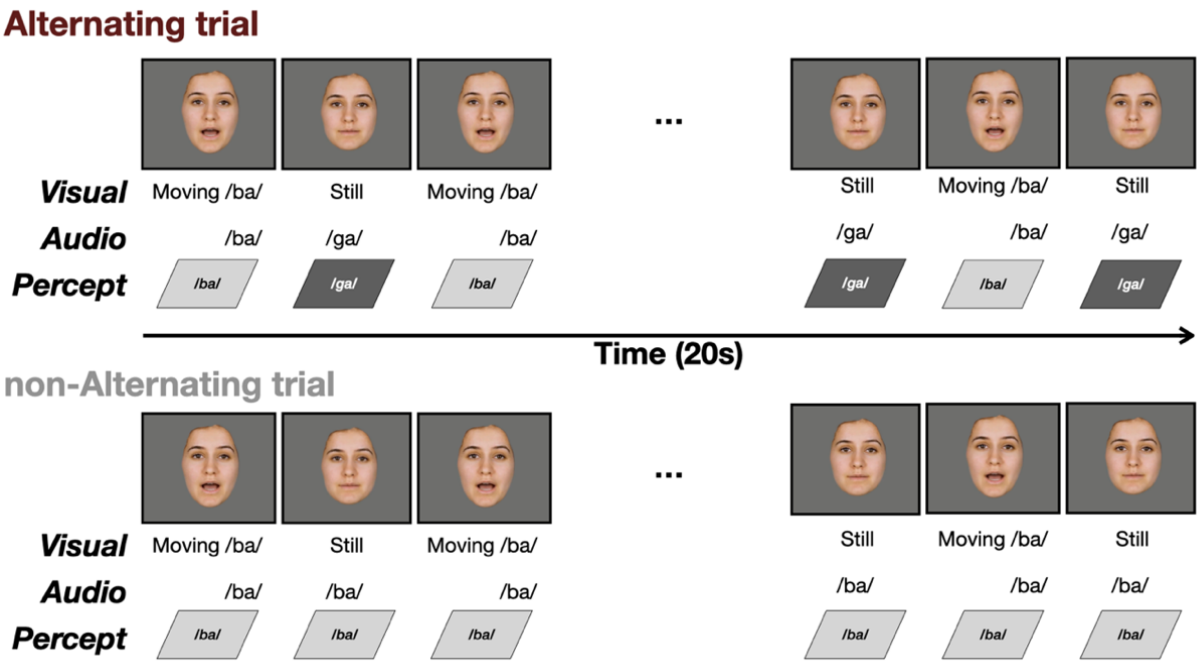


Figure 12. *Schematic illustration of the experimental procedure in the Alternating and non-Alternating trials.* Only one of the eight White stimulus faces, the visual syllable /ba/, and two audio syllables (/ba/ and /ga/) are presented here for illustration purposes.

The trials were played to infants in a randomized order, and no identical trial was presented twice in a row. The experiment terminated after infants finished all eight trials or if they stopped looking to the screen. Face-race conditions were counterbalanced across all infants.

Results and Discussion.

Infants preferred alternating over non-alternating auditory patterns.

On average, the participants finished 7.9 trials. To examine whether infants exhibited a preference for auditory sequences that alternate regularly between two syllables versus those that repeated a single syllable, we performed a paired-sample *t*-test comparing all infants' average looking time between Alternating and non-Alternating trials. As Figure 13 shows, the analysis revealed that infants looked significantly longer to the Alternating (M = 14.72 s, SD = 3.23) than

non-Alternating trials (M = 13.74 s, SD = 3.02; $t(22) = 2.29$, $p = .032$, Cohen's $d = 0.48$), indicating a reliable preference for regularly alternating auditory sequences over repetitive ones.

Prior to the analysis, we confirmed that the difference scores between two trial types—calculated by subtracting infants' total looking time in the non-Alternating trials from that in the Alternating trials—were normally distributed, as indicated by a Shapiro–Wilk test ($W = 0.93$, $p = .120$) and visual inspections of the histogram and Q-Q plot. These results all supported the normality assumption, further justifying the appropriateness of a paired-sample t-test.

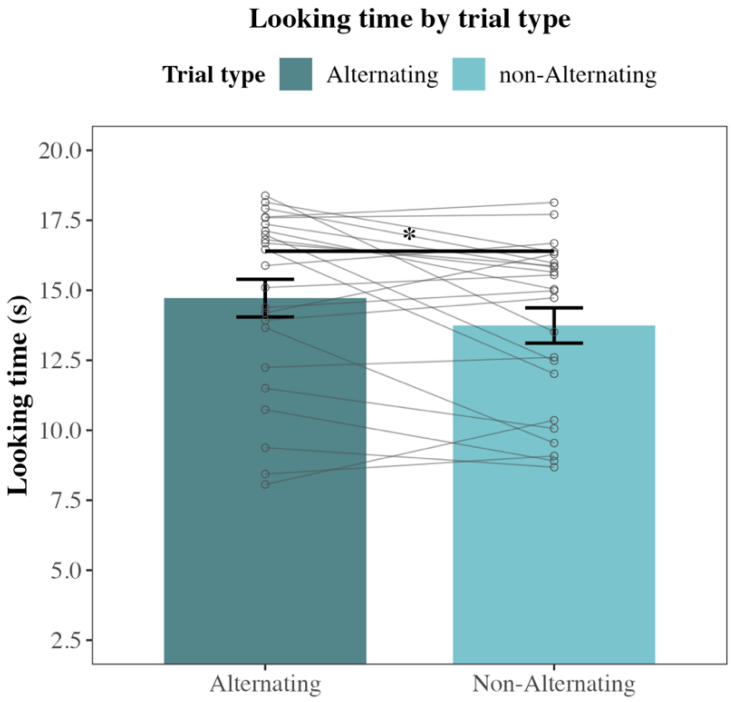


Figure 13. *Mean looking times for the Alternating and non-Alternating trials in Experiment 3.* The asterisks represent the statistically significant difference in looking time between two types of trials. Error bars represent ±1 standard error from the mean. Each dot represents an individual participant's looking time in each trial type, and the lines connecting paired dots reflect within-participant comparisons.

This preference was stable across infancy and independent of face-race.

To investigate whether this auditory preference changed with age, we first conducted a Pearson's correlation test between infants' age and their looking time difference between the two trial types. The difference between trial looking time (i.e., trial type preference scores) was calculated using the formula below:

$$\text{Trial preference score} = \text{Total looking time}_{\text{Alter}} - \text{Total looking time}_{\text{non-Alter}}$$

This test allowed us to examine whether older infants displayed a different preference for structured, alternating auditory patterns compared to younger infants. As shown in Figure 14, the analysis revealed no significant age-related trend ($r(21) = 0.14$, $p = .521$), suggesting that the preference for regularly alternating sound sequence was stable across the latter half of the first year in life.

Before the Pearson's correlation analysis, we confirmed that the linearity and homoscedasticity assumptions of for the Pearson's *r* were met, and the residuals from the linear model (*trial preference score ~ ageDays*) were normally distributed ($W = 0.939$, $p = .170$). Cook's distance analysis identified no potentially influential observation.

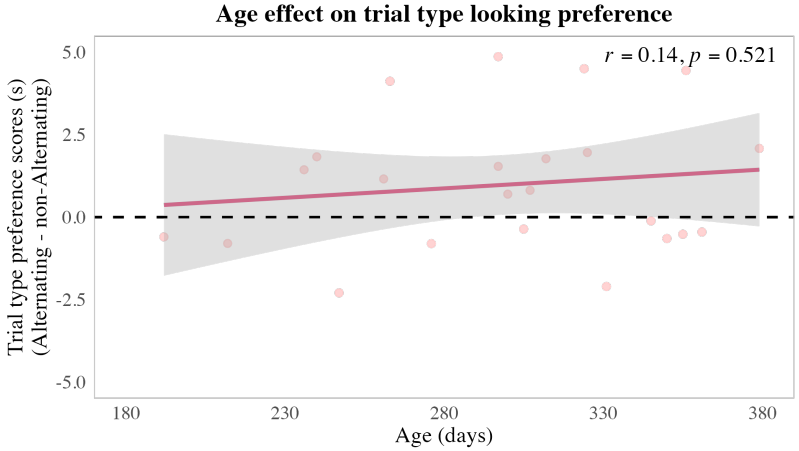


Figure 14. *Age-related change in the trial type looking preference collapsing all infants' data (own-race and other-race face conditions) in Experiment 3.* Each dot represents data from an

individual participant. The horizontal dashed line indicates equal looking to Alternating and non-Alternating trials. Dots above and below the dashed line are individuals who looked longer to the Alternating and non-Alternating trials, respectively. The shaded area around the regression line represents the confidence interval of the regression model.

While the Pearson correlation results provided a general overview of the relationship between age and infants' preference for Alternating versus non-Alternating trials, it did not account for the potential moderating effects of the races of the faces presenting the audiovisual speech stimuli. To address this, we conducted a multiple linear regression analysis predicting infants' trial type preference scores from face-race condition (own- vs. other-race), with age in days included as a continuous covariate.

The model revealed that neither face-race condition nor age significantly predicted infants' trial type preference. Specifically, the effect of face condition was not significant ($b = –0.10$, $SE = 0.48$, $t(20) = –0.21$, $p = .839$; partial $R^2 = .002$), and age in days also did not significantly predict infants' looking preference ($b = 0.01$, $SE = 0.01$, $t(20) = 0.58$, $p = .570$; partial $R^2 = .016$). The overall model was not significant ($F(2, 20) = 0.22$, $p = .800$), explaining minimal variance in the outcome ($R^2 = .022$, *Adjusted* $R^2 = –.076$). These results suggest that infants' preference for the Alternating over non-Alternating trials was not systematically influenced by either the race of the faces or the infants' age. In other words, infants' preference for Alternating trials remained consistent across 6 to 12 months of age and was unaffected by whether the audiovisual information was enacted by an own-race or other-race face.

To further illustrate these patterns, we conducted exploratory Pearson correlation analyses within each race condition. In the own-race condition, no significant correlation was found between age and Alternating trial preference ($r = –.13$, $p = .760$), suggesting that infants'

preference for regularly alternating over repetitive auditory patterns remained stable across the latter half of the first year. Similarly, in the other-race condition, the correlation between age and the preference for Alternating trials was also not significant (r = .23, p = .408). Although modest numerical differences were observed across face-race groups, the wide confidence intervals (own-race: [–.76, .63]; other-race: [–.31, .66]) suggest considerable uncertainty in the effect sizes. This reinforces the interpretation that any potential influence of face-race familiarity on infants' preference patterns is likely minimal and should be dealt with caution. Together, these parallel patterns across the two race conditions support the findings from the multiple linear regression analysis, that neither face-race familiarity nor infants' age modulated their preference for alternating over non-alternating sound patterns. Figure 15 plots the relationships between age and Alternating trials looking preference separately for each face-race condition.

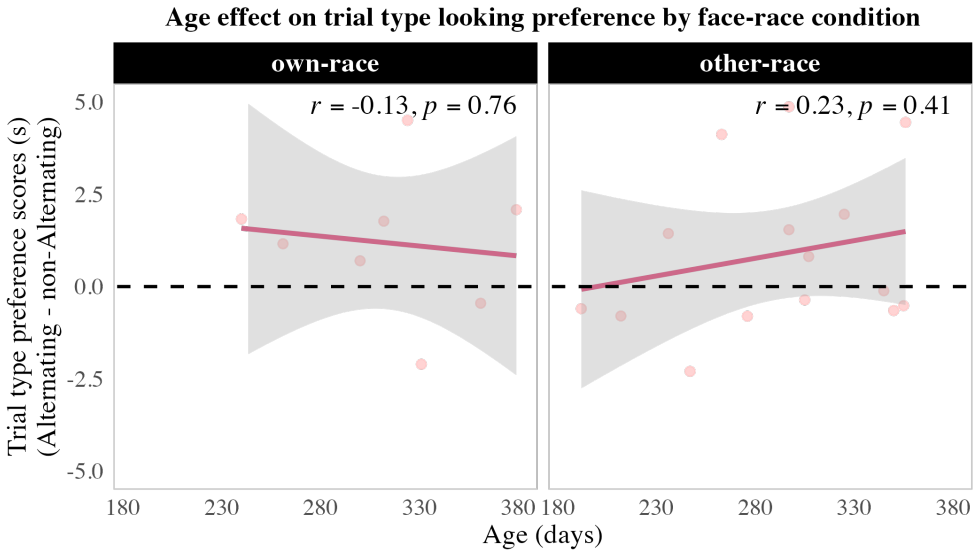


Figure 15. *Age-related changes in trial type looking preferences between the Alternating and non-Alternating trials by face-race condition in Experiment 3.* Age-related trend in the preferences between the two trial types by own-race (left panel) and other-race (right panel) conditions. Individual dots represent data from each participant. The horizontal dashed baseline

indicates an equal looking time in the two trial types. Any dot above and below the baseline represents longer looking time to the Alternating and non-Alternating trials, respectively. The shaded area around the trending line represents the confidence interval of the regression model.

In sum, using an audiovisual speech stimulus presentation similar to Experiments 1 and 2—with auditory and visual speech syllables being always congruent—Experiment 3 revealed two primary findings that support and clarify those from earlier experiments. First, infants manifested a pronounced and robust preference for regularly alternating over repetitive sound pattern across the latter half of the first year of life, the age range tested in this study. This finding provides theoretical support for our explanatory account of infants' looking preferences in response to the McGurk and non-McGurk trials and their corresponding levels of audiovisual speech integration (i.e., if infants perceived a regular auditory alternation, they would prefer it over the repetitive auditory pattern, suggesting a sufficiently developed integration capacity that enabled them to perceive a fused auditory percept). Second, this preference was not modulated or impaired the race of the stimulus faces. In other words, the other-race effect does not occur when the auditory and visual speech cues were always matched, highlighting the possibility that ORE may only emerge when bimodal speech information is conflicting yet interpretable within a shared representational and perceptual space.

General Discussion

Through a set of three experiments, this study investigated the capacity for audiovisual speech integration during the second half of the beginning year of life, focusing on two primary research questions. First, we examined whether infants demonstrate the ability to integrate auditory and visual speech cues and, if so, whether this capacity strengthens between 6 and 12

months of age (i.e., the 'differentiation' account of progressive multisensory development from a Gibsonian perspective)—a crucial developmental window marked by the emergence of endogenous attention and increased coordination between speech perception and production, both of which contribute to heightened visual attention to articulatory mouth movements. Second, we examined whether this integration is susceptible to interference from faces of other races during the same developmental period (i.e., the perceptual tuning account of regressive multisensory development)—a time that coincides with the emergence of the other-race effect (ORE) in unimodal face processing. If such interference were observed, we further explored whether differential face-scanning strategies might account for the ORE, with particular focus on infants' allocation of attention to the mouth region.

To address these questions, we applied the McGurk effect (McGurk & MacDonald, 1976) as a proxy for audiovisual speech integration within a perception-based behavioural task modeled on the Stimulus-Alternation Preference Procedure paradigm (SAPP; Best & Jones, 1988). Specifically, we relied on infants' differential looking times to two types of experimental trials—McGurk (Aba-Vga and Apa-Vka, typically fusible into /da/ and /ta/, respectively) and non-McGurk (Aga-Vba and Aka-Vpa, non-fusible and thus likely remained faithful to auditory /ga/ and /ka/)—as indicators of audiovisual speech integration. In addition, we enacted these trials with own-race (White; Experiment 1) or other-race (East-Asian; Experiment 2) faces to test whether the integration strength is modulated by face-race familiarity, thereby examining the potential influence of ORE on audiovisual speech integration in infancy.

Regarding our first research question, we hypothesized that infants' audiovisual speech integration would be present but still developing within the tested age range. Partially confirming this hypothesis, analysis of eye-tracking data from the 6- to 12-month-old infants presented with

own-race faces (Experiment 1) revealed evidence of a robust integration effect that remained stable across this developmental window. As reflected in significantly longer overall looking times to McGurk trials compared to non-McGurk trials, these findings suggest that infants may have perceived a regularly alternating auditory pattern, supported by a mature cross-modal speech integration system that reliably generated the fused percept each time auditory and visual syllables were paired in the moving-still-moving-still face presentation. This perceptual alternation likely sustained their attention longer than the perceptually repetitive non-McGurk trials—a preference directly corroborated in Experiment 3 using alternating and alternating trials with congruent audiovisual syllables. These results complement previous work using the McGurk effect to examine early audiovisual speech integration. Behavioural evidence from Burnham and Dodd (2004) demonstrated that 4.5-month-olds perceive a fused /da/ or /tha/ when exposed to Aba-Vga stimuli, and Kushnerenko et al. (2008) showed, via electrophysiological recordings, that 5-month-olds assimilate Aba-Vga into a legally integrated percept. Our findings further extend this prior work by introducing an additional integrable pair (Apa-Vka) and presenting these stimuli across eight different faces, thereby enhancing the generalizability of early and robust audiovisual speech integration in infancy.

We propose two possible accounts for the present finding that White Canadian infants already demonstrate a well-developed capacity for cross-modal speech integration capacity by the time of their participation (i.e., 6 months). First, as reviewed in *Introduction*, infants' innate structural sensitivity to the bimodal representation of speech may enable an early capacity to encode auditory and visual speech cues within a shared perceptual space. Specifically, from birth, infants exhibit strong phonetic matching abilities for both native and non-native vowels (Albridge et al., 1999; Kuhl & Meltzoff, 1982, 1984, 1988; Walton and Bower, 1993; Patterson

& Werker, 1999, 2003) as well as consonants (Danielson et al., 2017; MacKain et al., 1983), demonstrating a remarkably early sensitivity to the auditory and visual representations of phonemes. This built-in phonological knowledge may equip infants with the prerequisite ability to detect correspondences and compatibilities between visual articulations (i.e., mouth shape) and the auditory speech sounds. Given that the McGurk effect primarily arises from the visual resemblance between the mouth shape of the presented visual syllable and the fused percept as well as the auditory similarities between the physical auditory syllable and the percept (e.g., Lindborg et al., 2021; Tiippana et al., 2023), infants may already be well-prepared to integrate McGurk-type syllable pairings even early in development. Further investigations are encouraged to explore the developmental onset of the McGurk effect in infancy to clarify whether this capacity shows strengthening trajectories before 6 months, which helps to evaluate the Gibsonian perspective of the differentiation account of multisensory development. Electrophysiological methods—such as EEG and ERP—may be particularly suitable for this purpose, given that the smooth-pursuit eye movements required for high-quality eye-tracking data depend on the maturation of the oculomotor system, which typically begins around 4 months of age (Phillips et al., 1997).

Another possibility is that the participants in Experiment 1 had arguably received sufficient and high-quality visual speech inputs by 6 months, which could facilitate an early and evident capacity to integrate visual speech cues with auditory perception. Visual speech exposure during early postnatal life have been shown to be crucial to the development of multisensory perception. Indeed, research on animals (Carlson et al., 1987; Wallace & Stein, 1997; Wallace et al., 2004) and humans (Putzar et al., 2007; Putzar et al., 2010) provides converging evidence that early visual deprivation leads to deficits in audiovisual integration later in life. These deprivation

effects suggest that attending to the visual source of speech is not merely beneficial but may be integral for the ongoing development of multisensory perceptual abilities during infancy. Although the current study did not directly assess infants' visual speech exposure, we propose that, within the context of Canada's extended parental leave policy (UNICEF, 2019), the infants likely experienced considerable exposure to coordinated audiovisual speech inputs by 6 months, supporting early integration development. Parents' responses to the post-session questionnaire further support this possibility, indicating an average of 33.5 hours spent daily with both mothers and fathers combined (see Parental Questionnaire Data in *Supplementary Material*s for details).

Beyond the quantity of exposure, we argue that the quality of audiovisual speech inputs was likely high and conducive to infants' attention and learning of the coordination between heard and seen speech. Indeed, a substantial body of work has shown that caregivers adapt their communicative behaviours in interactions with infants—for instance, through temporally contingent feedback on infant vocalizations (Elmlinger et al., 2019; Goldstein & Schwade, 2008), increased frequency, exaggerated prosodic features, and expanded vowel spaces in infant-directed speech (Fernald et al., 1989; Liu et al., 2003; Nencheva & Lew-Williams, 2022; Soderstrom et al., 2021), etc. These heightened linguistic properties can highlight the integrative nature of auditory and visual speech signals, thereby fostering early cross-modal integration. Future research could explore potential links between the duration (i.e., total time) and quality (i.e., the richness of infant-directed communication; see Kosie & Lew-Williams, 2024, for a new multidimensional framework) of caregiver-infant engagement and the development of multisensory perception, including—but not limited to—speech-related integration abilities.

In addressing our second research question, we hypothesized that other-race faces would disrupt audiovisual speech integration, resulting in an attenuated McGurk effect that would show

no age-related increase—or possibly even a decline. Supporting this hypothesis, eye-tracking data from 6- to 12-month-old White Canadian infants viewing other-race faces (Experiment 2) provided evidence of a less-developed audiovisual speech integration capacity that remained unchanged across the tested age range. While infants still showed signs of integration, the strength appeared diminished, as indicated by their significantly longer looking time to non-McGurk than to McGurk trials. This reversed looking preference suggests that infants may have perceived the McGurk trials as presentations of a sporadic and unpredictable alternation of sounds—a perceptual outcome potentially driven by an inconsistent integration process that rendered the fused percept only intermittently when auditory and visual syllables were paired. This irregularity may have led infants to disengage from the McGurk trials and instead direct their attention toward the more predictable structure of the non-McGurk trials—a preference that, incidentally, echoes the observed bias for regularly alternating over repetitive patterns in Experiment 3.

In line with previous research examining the McGurk effect with other-race faces (Ujiie et al., 2020, 2021), the current finding demonstrates an other-race effect (ORE) on audiovisual speech integration, suggesting a perceptual advantage for own-race faces in multisensory speech perception. Notably, our data show no age-related changes, indicating that the ORE is already evident by 6 months of age—consistent with the well-established developmental timeline of ORE in face perception (e.g., Kelly et al., 2007, 2009; Xiao et al., 2018). Importantly, this finding contributes to the growing body of work proposing that perceptual narrowing is a pan-sensory, modality-general process (Lewkowicz & Ghazanfar, 2009), extending beyond the commonly studied unimodal domains of speech perception (vowels: Bosch & Sebastián-Gallés, 2003; consonants: Rivera-Gaxiola et al., 2005; lexical tones: Mattock et al., 2008; visual speech:

Sebastián-Gallés et al., 2012; Weikum et al., 2007) and face recognition (species: Pascalis et al., 2002; Simpson et al., 2011; race: Kelly et al., 2007, 2009; age: Kobayashi et al., 2018; Macchi Cassia et al., 2013). Furthermore, by incorporating familiar- and unfamiliar-race faces, our findings complement and extend prior work suggesting that perceptual narrowing in audiovisual speech perception arises from the perceptual tuning of native language inputs. Specifically, infants appear to be unable to match auditory and visual speech cues in non-native languages, whether at the syllabic level (Pons et al., 2009) or in fluent speech (Kubicek et al., 2014a, 2014b), as a consequence of reduced sensitivity to non-native auditory contrasts. Taken together, these findings underscore the mutually influential roles of auditory and visual experience, suggesting that narrowing in either modality can interfere with successful audiovisual speech perception. This supports the view that perceptual narrowing for both faces and languages may unfold into a cross-modal, domain-general process during the latter half of the first year of life.

Interestingly, we did not observe any other-race effect (ORE) on infants' perception of congruent audiovisual syllable pairings in Experiment 3. Regardless of whether the faces were of own- or other-race, infants consistently preferred alternating over non-alternating audiovisual syllable sequence—a robust and stable pattern across the tested age range. One possible explanation for the absence of an interfering ORE in this context lies in the differing attentional demands across tasks used in Experiments 2 and 3. In essence, Experiment 3 asked infants to differentiate between repetitive versus alternating auditory patterns, with visual cues that were always congruent with the auditory cues, thus providing redundant information; in contrast, Experiment 2 required infants to resolve a cross-modal conflict between mismatched auditory and visual syllables to generate a fused percept—a task that arguably demands greater attentional load. If other-race faces functions as a visual distractor (i.e., as an unfamiliar social category)

across both tasks, their disruptive impact would be expected to more pronounced in, and detrimental to, the integration task (Experiment 2), where successful performance hinges on coordinating inputs across modalities. This interpretation aligns with prior findings that high attentional load—especially from competing visual information—can reduce the incidence of McGurk fusion in adults (e.g., Alsius et al., 2005; Tiippana et al., 2004). Thus, the absence of an ORE in Experiment 3 does not contradict our broader conclusion that racially unfamiliar faces can impair audiovisual speech integration. Rather, it highlights how task demands moderate the manifestation of the ORE, with integration-based tasks being more susceptible to disruption than congruence-based ones.

Unlike previous findings that reported a complete absence of audiovisual integration when infants were presented with other-race faces (Ujiie et al., 2020, 2021), our results suggest that integration capacity is impaired but not entirely absent. We propose two possible explanations for this discrepancy, considering both methodological and environmental factors. First, the previous studies employed a memory-based familiarization paradigm, in which infants were familiarized with Apa-Vka stimuli and later tested with /pa/, the physical auditory syllable. This approach may have failed to capture infants' real-time integration. Specifically, infants may have perceived the fused percept during familiarization but, due to weakened recognition memory for other-race faces (Liu et al., 2011), they may have been unable to retain it through the test phase, preventing a novelty preference from emerging. In addition, the same study (Liu et al., 2011) found that infants allocated less visual attention to the internal features of other-race faces. Thus, a familiarization paradigm involving face-voice pairings followed by auditory-only testing may be insufficiently sensitive to detect weak or emerging integration. In contrast, the perception-based behavioral task used in our study, which presented continuous face-voice

pairings throughout the experimental session, may have enabled more sensitive detection by reducing reliance on memory and offering sustained exposure.

Another possible explanatory factor lies in the differing racial landscapes of the environments in which the infants were raised. The participants in Ujiie et al. (2020, 2021) were from Japan—a racially homogeneous society where infants have negligible visual exposure to other-race faces—potentially making the disruptive effect of other-race faces more pronounced than in our participants, who were raised in the racially diverse Greater Toronto Area (GTA). Admittedly, the heterogeneity of experimental paradigms complicates direct comparisons. However, using identical experimental setups and the same perception-based behavioral task paired with eye-tracking, the current study's participants were included in a cross-cultural comparison of the McGurk effect with Mandarin-speaking Chinese infants from a racially homogeneous city (see Yan et al., 2025). Similar to Japanese infants in Ujiie and colleagues' studies, Chinese infants showed no evidence of audiovisual speech integration, displaying comparable looking times to McGurk and non-McGurk trials. This comparison more directly underscores the role of postnatal visual experience in infants' immediate environments.

Although questionnaire data indicated minimal direct exposure to East-Asian faces among infants in our other-race condition (after race-based exclusion)—with only one infant having an East-Asian father and no other caregivers reporting regular weekly interactions with East-Asian individuals—prior research suggests that even brief (~8-minute) daily exposure to East-Asian faces over the course of three weeks suffice to restore White infants' sensitivity to these faces after a perceptual narrowing window (see Anzures et al., 2012, for evidence in 8- to 10-month-olds). This raises the possibility that, even without caregivers' conscious awareness, infants encounter and encode multisensory cues related to other-race faces in everyday settings,

such as during outings or restaurant visits. Supporting this idea, evidence shows that by 11 months, infants may already be sensitive to the racial and ethnic composition of their broader social environment (Singarajah et al., 2017) and can form expectations anchored in a speaker's racial identity (see May et al., 2019, for evidence of infants' flexible associations between language and race in a multilingual, multicultural metropolitan context). To build on these findings, future work could examine how many unfamiliar-race face exemplars are needed to mitigate the ORE in infants' audiovisual speech integration—that is, how much exposure to other-race faces would be sufficient to elicit a robust integration response comparable to that observed with own-race faces.

One important consideration in interpreting our findings is how faces of familiar versus unfamiliar races influenced infants' McGurk effect, which we explored using area-of-interest (AOI) analyses of their face-scanning patterns. Using a standardized face template (Xiao & Lee, 2018), our analysis of infants' proportional looking time to the mouth region revealed no significant differences when they viewed own- and other-race speakers, suggesting that variation in mouth-looking likely did not account for the observed ORE in audiovisual speech integration. Moreover, contrary to prior findings that increased fixation on a speaker's mouth supports stronger audiovisual integration—either from adults' eye-tracking data (Gurler et al., 2015) or infants' brain responses to audiovisual mismatches (AVMMR; Kushnerenko et al., 2013), both using McGurk syllables—we found no such association in our data. Interestingly, we observed a (marginally significant) trend in the opposite direction: infants who exhibited stronger McGurk fusion tended to look proportionally less at the mouth, raising questions for future work about whether face-race familiarity modulates infants' mouth-looking behaviours. Several factors may help explain this null effect of mouth-looking on the observed ORE, including the lack of rich

semantic content in our stimuli and the use of stimulus faces that were not life-sized, both of which may have reduced the precision of our mouth-looking measurements. The former limitation may also explain infants' consistent preference for the eyes over the mouth in our task—a pattern that contrasts with findings using more naturalistic speech stimuli (e.g., Lewkowicz & Hansen-Tift, 2012; Pons et al., 2015).

Although mouth-looking alone fails to convincingly explain the ORE on audiovisual speech integration, we propose the possibility that face-race familiarity differentially modulates infants' real-time visual attention across the experimental session in ways that impaired or facilitated integration. Specifically, face-race familiarity appears to shape infants' dynamic adjustments of attention on the eyes and mouths—AOIs central to social pragmatic cues (i.e., non-verbal communicative signals that help infants interpret intentions) and articulatory speech cues—over the experimental session. When viewing other-race speakers, infants demonstrated a significant bias toward the eyes starting early on (i.e., Block 2) before any signs of habituation to the mouth (i.e., starting to emerge by Block 3). Such a reduced prioritization of articulatory speech cues—whether voluntarily motivated to explore communicative intents or establish social contacts with someone who is visually salient to come from a different community membership (Uttley et al., 2013; Weatherhead & Werker, 2022; Yuan et al., 2019), or compulsory as a default turn-to upon failure to efficiently extract visual speech information (Ujiie & Takahashi, 2022)—arguably attenuate the robustness of integration. These interpretations build on evidence suggesting that selective attention to the mouth over eyes supports infants' exploitation of audiovisual speech cues (e.g., Lewkowicz & Hansen-Tift, 2012, Pons et al., 2015).

However, infants viewing own-race faces only began to show a notable visual bias toward the eyes over the mouth in Block 3—one block later than infants in the other-race

condition—and only after exhibiting a clear habituation-like response to articulatory speech information, suggesting that the stimuli had become less informative or more predictable. We reason that, whether due to a delayed onset or a strategic shift following sufficient exposure, these infants may have afforded themselves more opportunities to extract visual speech cues and integrate them with auditory input. This contrast in infants' scanning strategies for own-race versus other-race faces suggests that they deploy different patterns of visual attention depending on facial familiarity (Liu et al., 2011; Wheeler et al., 2011)—potentially prioritizing social communicative intent over speech-relevant information with other-race faces, but not with own-race faces. Such differential prioritization may help explain the race-dependent differences observed in McGurk perception.

In summary, the current study provides evidence for the early emergence of a robust capacity for audiovisual speech integration—one that is shaped by infants' immediate social environments. Specifically, infants readily incorporate visual speech cues from own-race speakers but show impaired integration when the same information is presented by other-race speakers, indicating that audiovisual speech perception is susceptible to the other-race effect (ORE). These findings highlight the role of postnatal experience with faces in shaping audiovisual speech perception and contribute to the growing body of research suggesting that perceptual narrowing extends beyond unimodal face and speech processing to operate at a cross-modal, pan-sensory level. From the perspective of multisensory development, the present study supports the perceptual tuning account, shedding light on how infants navigate the multisensory world through both regressive and progressive developmental processes.

Conclusion

The second half of the first year is a crucial developmental period: the perceptual system becomes rapidly tuned to the ecologically relevant social and linguistic cues, the vocal repertoire matures to produce speech-like babbling, and endogenous attention emerges to support voluntary visual exploration. These developments raise a fundamental question: how ready is the early multisensory system to integrate auditory and visual speech cues—an ability that lays the foundation for language acquisition and later socio-cognitive and socio-emotional development? Using the McGurk effect as an index of visual influence on auditory speech perception, the present study demonstrates that infants have already developed a robust capacity to integrate speech-related audiovisual signals from own-race speakers by six months of age—a capacity that remains stable across the latter half of the beginning year of life. Moreover, this integration appears to be shaped by infants' visual experience with own-race faces in their immediate social environment, as their integration is evidently interfered by other-race faces by six months—a timing that coincides with the onset of perceptual narrowing in unimodal face recognition. We propose that, as a salient indicator of a different community membership, other-race faces prompt infants to gravitate toward social-pragmatic (e.g., communicative intent) over articulatory speech cues. Taken together, these results contribute to our understanding that (1) speech is presented bimodally in infancy and processed in an integrative manner, (2) perceptual tuning of early visual experience shapes audiovisual integration, suggesting that perceptual narrowing is a cross-modal, pan-sensory phenomenon, (3) multisensory perceptual development has a regressive nature before/while proceeding progressively, and (4) face race, as a salient social category, may modulate infants' prioritization of facial information.

References

Al Roumi, F., Planton, S., Wang, L., & Dehaene, S. (2023). Brain-imaging evidence for compression of binary sound sequences in human memory. *ELife, 12*, e84376. https://doi.org/10.7554/eLife.84376

Aldridge, M. A., Braga, E. S., Walton, G. E., & Bower, T. G. R. (1999). The intermodal representation of speech in newborns. *Developmental Science*, *2*(1), 42–46. https://doi.org/10.1111/1467-7687.00052

Alsius, A., Navarra, J., Campbell, R., & Soto-Faraco, S. (2005). Audiovisual integration of speech falters under high attention demands. *Current biology*, *15*(9), 839-843. 10.1016/j.cub.2005.03.046

Altvater-Mackensen, N., Jessen, S., & Grossmann, T. (2017). Brain responses reveal that infants' face discrimination is guided by statistical learning from distributional information. *Dev Sci*, *20*(2). https://doi.org/10.1111/desc.12393

Anzures, G., Wheeler, A., Quinn, P. C., Pascalis, O., Slater, A. M., Heron-Delaney, M., ... & Lee, K. (2012). Brief daily exposures to Asian females reverses perceptual narrowing for Asian faces in Caucasian infants. *Journal of experimental child psychology*, *112*(4), 484-495. https://doi.org/10.1016/j.jecp.2012.04.005

Bahrick, L. E., & Hollich, G. J. (2020). Intermodal perception. In J. B. Benson (Ed.), *Encyclopedia of infant and early childhood development* (2nd ed., pp. 202–217). Elsevier. https://doi.org/10.1016/B978-0-12-809324-5.23594-3

Bahrick, L. E., & Lickliter, R. (2002). Intersensory redundancy guides early perceptual and cognitive development. Advances in Child Development and Behavior, 30, 153–187. https://doi.org/10.1016/s0065-2407(02)80041-6

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. https://doi.org/10.1111/j.0963-7214.2004.00283.x

Bahrick, L. E., Netto, D., & Hernandez-Keif, M. (1998). Intermodal perception of adult and child faces and voices by infants. *Child development*, *69*(5), 1263-1275. https://doi.org/10.1111/j.1467-8624.1998.tb06210.x

Bar-Haim, Y., Ziv, T., Lamy, D., & Hodes, R. M. (2006). Nature and nurture in own-race face processing. *Psychological science*, *17*(2), 159-163. https://doi.org/10.1111/j.1467-9280.2006.01679.x

Beauchamp, M. S., Argall, B. D., Bodurka, J., Duyn, J. H., & Martin, A. (2004). Unraveling multisensory integration: patchy organization within human STS multisensory cortex. *Nature neuroscience*, *7*(11), 1190-1192.10.1038/nn1333

Bergelson, E. (2020). The comprehension boost in early word learning: Older infants are better learners. *Child development perspectives*, *14*(3), 142-149. https://doi.org/10.1111/cdep.12373

Best, C., & Jones, C. (1998). Stimulus-alternation preference procedure to test infant speech discrimination. *Infant Behavior and Development, 21*, 295. https://doi.org/10.1016/S0163-6383(98)91508-9

Birch, H. G., & Lefford, A. (1967). Visual differentiation, intersensory integration, and voluntary motor control. *Monographs of the society for research in child development*, *32*(2), 1-87. https://doi.org/10.2307/1165792

Birulés, J., Bosch, L., Brieke, R., Pons, F., & Lewkowicz, D. J. (2019). Inside bilingualism: Language background modulates selective attention to a talker's mouth. *Developmental science*, *22*(3), e12755. https://doi.org/10.1111/desc.12755

Bosch, L., & Sebastián-Gallés, N. (2003). Simultaneous bilingualism and the perception of a language-specific vowel contrast in the first year of life. *Language and speech*, *46*(2-3), 217-243. https://doi.org/10.1177/00238309030460020801

Bothwell, R. K., Brigham, J. C., & Malpass, R. S. (1989). Cross-racial identification. *Personality and Social Psychology Bulletin*, *15*(1), 19-25. https://doi.org/10.1177/0146167289151002

Bovo, R., Ciorba, A., Prosser, S., & Martini, A. (2009). The McGurk phenomenon in Italian listeners. *Acta Otorhinolaryngologica Italica, 29*(4), 203. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2816368/

Bruce, V., & Young, A. (1986). Understanding face recognition. *British journal of psychology*, *77*(3), 305-327. https://doi.org/10.1111/j.2044-8295.1986.tb02199.x

Burnham, D., & Dodd, B. (1996). Auditory-visual speech perception as a direct process: The McGurk effect in infants and across languages. In D. G. Stork & M. E. Hennecke (Eds.), *Speechreading by humans and machines* (Vol. 150, pp. 103–114). Springer. https://doi.org/10.1007/978-3-662-13015-5_7

Burnham, D., & Dodd, B. (2004). Auditory–visual speech integration by prelinguistic infants: Perception of an emergent consonant in the McGurk effect. *Developmental Psychobiology, 45*(4), 204–220. https://doi.org/10.1002/dev.20032

Burnham, D., & Lau, S. (1998). The effect of tonal information on auditory reliance in the McGurk effect. In *Proceedings of the Auditory-Visual Speech Processing* (*AVSP) Conference* (pp. 37-42).

Calvert, G. A., Campbell, R., & Brammer, M. J. (2000). Evidence from functional magnetic resonance imaging of crossmodal binding in the human heteromodal cortex. *Current biology*, *10*(11), 649-657. 10.1016/s0960-9822(00)00513-3

Carlson, S., Pertovaara, A., & Tanila, H. (1987). Late effects of early binocular visual deprivation on the function of Brodmann's area 7 of monkeys (Macaca arctoides). *Developmental Brain Research*, *33*(1), 101-111. https://doi.org/10.1016/0165-3806(87)90180-5

Celce-Murcia, M., Brinton, D. M., & Goodwin, J. M. (2010). *Teaching pronunciation hardback with audio CDs (2): A course book and reference guide*. Cambridge University Press.

Chandrasekaran, C., Trubanova, A., Stillittano, S., Caplier, A., & Ghazanfar, A. A. (2009). The natural statistics of audiovisual speech. *PLoS computational biology*, *5*(7), e1000436. https://doi.org/10.1371/journal.pcbi.1000436

Colombo, J. (2001). The development of visual attention in infancy. *Annual review of psychology*, *52*(1), 337-367. https://doi.org/10.1146/annurev.psych.52.1.337

Danielson, D. K., Bruderer, A. G., Kandhadai, P., Vatikiotis-Bateson, E., & Werker, J. F. (2017). The organization and reorganization of audiovisual speech perception in the first year of life. *Cognitive Development*, *42*, 37-48. https://doi.org/10.1016/j.cogdev.2017.02.004

Desjardins, R. N., & Werker, J. F. (2004). Is the integration of heard and seen speech mandatory for infants? *Developmental Psychobiology*, *45*(4), 187–203. https://doi.org/10.1002/dev.20033

Desjardins, R. N., Rogers, J., & Werker, J. F. (1997). An exploration of why preschoolers perform differently than do adults in audiovisual speech perception tasks. *Journal of Experimental Child Psychology*, *66*(1), 85–110. https://doi.org/10.1006/jecp.1997.2379

Elmlinger, S. L., Schwade, J. A., & Goldstein, M. H. (2019). The ecology of prelinguistic vocal learning: Parents simplify the structure of their speech in response to babbling. *Journal of child language*, *46*(5), 998-1011. https://doi.org/10.1017/S0305000919000291

Farroni, T., Johnson, M. H., Menon, E., Zulian, L., Faraguna, D., & Csibra, G. (2005). Newborns' preference for face-relevant stimuli: Effects of contrast polarity. *Proceedings of the National Academy of Sciences*, *102*(47), 17245-17250. https://doi.org/10.1073/pnas.0502205102

Fernald A, Taeschner T, Dunn J, Papousek M, de Boysson-Bardies B, Fukui I. A cross-language study of prosodic modifications in mothers' and fathers' speech to preverbal infants. *Journal of Child Language*. 1989;16(3):477-501. 10.1017/S0305000900010679

Fuster Duran, A. (1995). McGurk effect in Spanish and German listeners: Influences of visual cues in the perception of Spanish and German conflicting audio-visual stimuli. In *Proceedings of Eurospeech 1995* (pp. 295–298).

Gibson, E. J. (1984). Perceptual development from the ecological approach. *Advances in developmental psychology*, *3*, 243-286.

Gijbels, L., Lee, A. K., & Lalonde, K. (2025). Integration of audiovisual speech perception: From infancy to older adults. *The Journal of the Acoustical Society of America*, *157*(3), 1981-2000. https://doi.org/10.1121/10.0036137

Goldstein, M. H., & Schwade, J. A. (2008). Social feedback to infants' babbling facilitates rapid phonological learning. *Psychological science*, *19*(5), 515-523. https://doi.org/10.1111/j.1467-9280.2008.02117.x

Guiraud, J. A., Tomalski, P., Kushnerenko, E., Ribeiro, H., Davies, K., Charman, T., ... & BASIS team. (2012). Atypical audiovisual speech integration in infants at risk for autism. *PloS one*, *7*(5), e36428. https://doi.org/10.1371/journal.pone.0036428

Gurler, D., Doyle, N., Walker, E., Magnotti, J., & Beauchamp, M. (2015). A link between individual differences in multisensory speech perception and eye movements. *Attention, Perception, & Psychophysics*, *77*(4), 1333–1341. https://doi.org/10.3758/s13414-014-0821-1

Hadjikhani, N., Hoge, R., Snyder, J., & de Gelder, B. (2008). Pointing with the eyes: the role of gaze in communicating danger. *Brain and cognition*, *68*(1), 1-8. https://doi.org/10.1016/j.bandc.2008.01.008

Hay, J. F., Graf Estes, K., Wang, T., & Saffran, J. R. (2015). From flexibility to constraint: The contrastive use of lexical tone in early word learning. *Child development*, *86*(1), 10-22. https://doi.org/10.1111/cdev.12269

Hockley, N. S., & Polka, L. (1994). A developmental study of audiovisual speech perception using the McGurk paradigm. *The Journal of the Acoustical Society of America*, *96*(5_Supplement), 3309–3309. https://doi.org/10.1121/1.410782

Hunnius, S., & Geuze, R. H. (2004). Developmental changes in visual scanning of dynamic faces and abstract stimuli in infants: A longitudinal study. *Infancy*, *6*(2), 231-255. https://doi.org/10.1207/s15327078in0602_5

Irwin, J. R., Whalen, D. H., & Fowler, C. A. (2006). A sex difference in visual influence on heard speech. *Perception & Psychophysics*, *68*(4), 582–592. https://doi.org/10.3758/BF03208760

Jusczyk, P. W., & Bertoncini, J. (1988). Viewing the development of speech perception as an innately guided learning process. *Language and Speech*, *31*(3), 217–238. https://doi.org/10.1177/002383098803100301

Kelly, D. J., Liu, S., Rodger, H., Miellet, S., Ge, L., & Caldara, R. (2011). Developing cultural differences in face processing: Developing cultural differences in face processing. *Developmental Science*, *14*(5), 1176–1184. https://doi.org/10.1111/j.1467-7687.2011.01067.x

Kelly, D. J., Quinn, P. C., Slater, A. M., Lee, K., Ge, L., & Pascalis, O. (2007). The other-race effect develops during infancy: Evidence of perceptual narrowing. *Psychological Science*, *18*(12), 1084–1089. https://doi.org/10.1111/j.1467-9280.2007.02029.x

Kidd, C., Piantadosi, S. T., & Aslin, R. N. (2012). The Goldilocks effect: Human infants allocate attention to visual sequences that are neither too simple nor too complex. *PLoS ONE*, *7*(5), e36399. https://doi.org/10.1371/journal.pone.0036399

Kidd, C., Piantadosi, S. T., & Aslin, R. N. (2014). The Goldilocks effect in infant auditory attention. *Child Development*, *85*(5), 1795–1804. https://doi.org/10.1111/cdev.12263

Kisilevsky, B. S., Hains, S. M. J., Brown, C. A., Lee, C. T., Cowperthwaite, B., Stutzman, S. S., Swansburg, M. L., Lee, K., Xie, X., Huang, H., Ye, H.-H., Zhang, K., & Wang, Z. (2009). Fetal sensitivity to properties of maternal speech and language. *Infant Behavior and Development*, *32*(1), 59–71. https://doi.org/10.1016/j.infbeh.2008.10.002

Kobayashi, M., Macchi Cassia, V., Kanazawa, S., Yamaguchi, M. K., & Kakigi, R. (2018). Perceptual narrowing towards adult faces is a cross-cultural phenomenon in infancy: A behavioral and near-infrared spectroscopy study with Japanese infants. *Developmental Science*, *21*(1). https://doi.org/10.1111/desc.12498

Kosie, J. E., & Lew-Williams, C. (2024). Infant-directed communication: Examining the many dimensions of everyday caregiver-infant interactions. *Developmental Science*, *27*(5), e13515. https://doi.org/10.1111/desc.13515

Kubicek, C., Gervain, J., De Boisferon, A. H., Pascalis, O., Lœvenbruck, H., & Schwarzer, G. (2014b). The influence of infant-directed speech on 12-month-olds' intersensory perception of fluent speech. *Infant Behavior and Development*, *37*(4), 644-651. https://doi.org/10.1016/j.infbeh.2014.08.010

Kubicek, C., Hillairet de Boisferon, A., Dupierrix, E., Pascalis, O., Lœvenbruck, H., Gervain, J., & Schwarzer, G. (2014a). Cross-modal matching of audio-visual German and French fluent speech in infancy. *PloS one*, *9*(2), e89275. https://doi.org/10.1371/journal.pone.0089275

Kuhl, P. K., & Meltzoff, A. N. (1982). The bimodal perception of speech in infancy. *Science*, *218*(4577), 1138-1141. 10.1126/science.7146899

Kuhl, P. K., & Meltzoff, A. N. (1984). The intermodal representation of speech in infants. Infant behavior and development, 7(3), 361-381. https://doi.org/10.1016/S0163-6383(84)80050-8

Kuhl, P. K., & Meltzoff, A. N. (1988). Speech as an intermodal object of perception. In A. Yonas (Ed.), *Perceptual development in infancy* (pp. 235–266). Lawrence Erlbaum Associates, Inc.

Kuhl, P. K., & Meltzoff, A. N. (1996). Infant vocalizations in response to speech: Vocal imitation and developmental change. *The journal of the Acoustical Society of America*, *100*(4), 2425-2438. https://doi.org/10.1121/1.417951

Kushnerenko, E., Teinonen, T., Volein, A., & Csibra, G. (2008). Electrophysiological evidence of illusory audiovisual speech percept in human infants. *Proceedings of the National Academy of Sciences*, *105*(32), 11442–11445. https://doi.org/10.1073/pnas.0804275105

Kushnerenko, E., Tomalski, P., Ballieux, H., Ribeiro, H., Potton, A., Axelsson, E. L., Murphy, E., & Moore, D. G. (2013). Brain responses to audiovisual speech mismatch in infants are associated with individual differences in looking behaviour. *European Journal of Neuroscience*, *38*(9), 3363–3369. https://doi.org/10.1111/ejn.12317

Lee, A.K.C., & Wallace, M.T. (2019). Visual influence on auditory perception. In: Lee, A., Wallace, M., Coffin, A., Popper, A., Fay, R. (Eds.), *Multisensory Processes*. Springer Handbook of Auditory Research, vol 68. Springer, Cham. https://doi.org/10.1007/978-3-030-10461-0_1

Legerstee, M. (1990). Infants use multimodal information to imitate speech sounds. *Infant behavior and development*, *13*(3), 343-354. https://doi.org/10.1016/0163-6383(90)90039-B

Lewkowicz, D. J. (2010). Infant perception of audio-visual speech synchrony. *Developmental psychology*, *46*(1), 66. https://doi.org/10.1037/a0015579

Lewkowicz, D. J., & Hansen-Tift, A. M. (2012). Infants deploy selective attention to the mouth of a talking face when learning speech. *Proceedings of the National Academy of Sciences*, *109*(5), 1431–1436. https://doi.org/10.1073/pnas.1114783109

Lewkowicz, D. J., & Kraebel, K. S. (2004). The value of multisensory redundancy in the development of intersensory perception. https://doi.org/10.7551/mitpress/3422.003.0049

Liu, H. M., Kuhl, P. K., & Tsao, F. M. (2003). An association between mothers' speech clarity and infants' speech discrimination skills. *Developmental science*, *6*(3), F1-F10. https://doi.org/10.1111/1467-7687.00275

Liu, S., Quinn, P. C., Wheeler, A., Xiao, N., Ge, L., & Lee, K. (2011). Similarity and difference in the processing of same- and other-race faces as revealed by eye tracking in 4- to 9-month-olds. *Journal of Experimental Child Psychology*, *108*(1), 180–189. https://doi.org/10.1016/j.jecp.2010.06.008

Lozano, I., Campos, R., & Belinchón, M. (2024). Sensitivity to temporal synchrony in audiovisual speech in early infancy: Current issues and future avenues. *Cognitive Development*, *70*, 101453. https://doi.org/10.1016/j.cogdev.2024.101453

Macchi Cassia, V., Kuefner, D., Picozzi, M., & Vescovo, E. (2009). Early experience predicts later plasticity for face processing: Evidence for the reactivation of dormant effects. *Psychological Science*, *20*(7), 853–859. https://doi.org/10.1111/j.1467-9280.2009.02376.x

MacDonald, J., & McGurk, H. (1978). Visual influences on speech perception processes. *Perception & psychophysics*, *24*(3), 253-257. https://doi.org/10.1177/016502547800100303

MacLin, O. H., & Malpass, R. S. (2001). Racial categorization of faces: The ambiguous race face effect. *Psychology, Public Policy, and Law*, *7*(1), 98.

Massaro, D. W., Cohen, M. M., & Smeele, P. M. (1996). Perception of asynchronous and conflicting visual and auditory speech. *The Journal of the Acoustical Society of America*, *100*(3), 1777-1786. https://doi.org/10.1121/1.417342

Massaro, D. W., Thompson, L. A., Barron, B., & Laren, E. (1986). Developmental changes in visual and auditory contributions to speech perception. *Journal of Experimental Child Psychology*, *41*(1), 93–113. https://doi.org/10.1016/0022-0965(86)90053-6

Mattock, K., Molnar, M., Polka, L., & Burnham, D. (2008). The developmental course of lexical tone perception in the first year of life. *Cognition*, *106*(3), 1367–1381. https://doi.org/10.1016/j.cognition.2007.07.002

Maurer, D., & Werker, J. F. (2014). Perceptual narrowing during infancy: A comparison of language and faces. *Developmental Psychobiology*, *56*(2), 154–178. https://doi.org/10.1002/dev.21177

May, L., Baron, A. S., & Werker, J. F. (2019). Who can speak that language? Eleven-month-old infants have language-dependent expectations regarding speaker ethnicity. *Developmental Psychobiology*, *61*(6), 859-873. https://doi.org/10.1002/dev.21851

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, *264*(5588), 746–748. https://doi.org/10.1038/264746a0

Nath, A. R., & Beauchamp, M. S. (2012). A neural basis for interindividual differences in the McGurk effect, a multisensory speech illusion. *Neuroimage*, *59*(1), 781-787. https://doi.org/10.1016/j.neuroimage.2011.07.024

Nencheva, M. L., & Lew-Williams, C. (2022). Understanding why infant-directed speech supports learning: A dynamic attention perspective. *Developmental Review*, *66*, 101047. https://doi.org/10.1016/j.dr.2022.101047

Oller, D. K. (2000). *The emergence of the speech capacity*. Psychology Press. https://doi.org/10.4324/9781410602565

Partanen, E., Kujala, T., Näätänen, R., Liitola, A., Sambeth, A., & Huotilainen, M. (2013). Learning-induced neural plasticity of speech processing before birth. *Proceedings of the National Academy of Sciences*, *110*(37), 15145–15150. https://doi.org/10.1073/pnas.1302159110

Pascalis, O., De Haan, M., & Nelson, C. A. (2002). Is face processing species-specific during the first year of life?. *Science*, *296*(5571), 1321-1323. 10.1126/science.1070223

Patterson, M. L., & Werker, J. F. (1999). Matching phonetic information in lips and voice is robust in 4.5-month-old infants. *Infant Behavior and Development*, *22*(2), 237-247. https://doi.org/10.1016/S0163-6383(99)00003-X

Patterson, M. L., & Werker, J. F. (2002). Infants' ability to match dynamic phonetic and gender information in the face and voice. *Journal of experimental child psychology*, *81*(1), 93-115. https://doi.org/10.1006/jecp.2001.2644

Patterson, M. L., & Werker, J. F. (2003). Two-month-old infants match phonetic information in lips and voice. *Developmental Science*, *6*(2), 191–196. https://doi.org/10.1111/1467-7687.00271

Phillips, J. O., Finocchio, D. V., Ong, L., & Fuchs, A. F. (1997). Smooth pursuit in 1-to 4-month-old human infants. *Vision Research*, *37*(21), 3009-3020. https://doi.org/10.1016/S0042-6989(97)00107-7

Piaget, J., & Cook, M. (1952). *The origins of intelligence in children* (Vol. 8, No. 5, pp. 18-1952). International Universities Press.

Pons, F., Bosch, L., & Lewkowicz, D. J. (2015). Bilingualism modulates infants' selective attention to the mouth of a talking face. *Psychological Science*, *26*(4), 490–498. https://doi.org/10.1177/0956797614568320

Putzar, L., Goerendt, I., Lange, K., Rösler, F., & Röder, B. (2007). Early visual deprivation impairs multisensory interactions in humans. *Nature neuroscience*, *10*(10), 1243-1245. 10.1038/nn1978

Putzar, L., Hötting, K., & Röder, B. (2010). Early visual deprivation affects the development of face recognition and of audio-visual speech perception. *Restorative neurology and neuroscience*, *28*(2), 251-257. https://doi.org/10.3233/RNN-2010-0526

Quinn, P. C., Yahr, J., Kuhn, A., Slater, A. M., & Pascalis, O. (2002). Representation of the gender of human faces by infants: A preference for female. *Perception*, *31*(9), 1109-1121. https://doi.org/10.1068/p3331

Richards, J. E., Reynolds, G. D., & Courage, M. L. (2010). The neural bases of infant attention. *Current Directions in Psychological Science*, *19*(1), 41-46. https://doi.org/10.1177/0963721409360003

Rivera-Gaxiola, M., Silva-Pereyra, J., & Kuhl, P. K. (2005). Brain potentials to native and non-native speech contrasts in 7-and 11-month-old American infants. *Developmental science*, *8*(2), 162-172. https://doi.org/10.1111/j.1467-7687.2005.00403.x

Rosenblum, L. D. (2008). Speech perception as a multimodal phenomenon. *Current Directions in Psychological Science*, *17*(6), 405–409. https://doi.org/10.1111/j.1467-8721.2008.00615.x

Rosenblum, L. D., Schmuckler, M. A., & Johnson, J. A. (1997). The McGurk effect in infants. *Perception and Psychophysics, 59*(3), 347–357. https://doi.org/10.3758/BF03211902

Saint-Amour, D., De Sanctis, P., Molholm, S., Ritter, W., & Foxe, J. J. (2007). Seeing voices: High-density electrical mapping and source-analysis of the multisensory mismatch negativity evoked during the McGurk illusion. *Neuropsychologia*, *45*(3), 587-597. https://doi.org/10.1016/j.neuropsychologia.2006.03.036

Schyns, P. G., Bonnar, L., & Gosselin, F. (2002). Show me the features! Understanding recognition from the use of visual information. *Psychological science*, *13*(5), 402-409. https://doi.org/10.1111/1467-9280.00472

Sebastián-Gallés, N., Albareda-Castellot, B., Weikum, W. M., & Werker, J. F. (2012). A bilingual advantage in visual language discrimination in infancy. *Psychological science*, *23*(9), 994-999. https://doi.org/10.1177/0956797612436817

Sekiyama, K. (1997). Cultural and linguistic factors in audiovisual speech processing: The McGurk effect in Chinese subjects. *Perception & Psychophysics*, *59*(1), 73–80. https://doi.org/10.3758/BF03206849

Sekiyama, K., & Burnham, D. (2008). Impact of language on development of auditory-visual speech perception. *Developmental science*, *11*(2), 306-320. https://doi.org/10.1111/j.1467-7687.2008.00677.x

Sekiyama, K., & Tohkura, Y. I. (1991). McGurk effect in non-English listeners: Few visual effects for Japanese subjects hearing Japanese syllables of high auditory intelligibility. *The Journal of the Acoustical Society of America*, *90*(4), 1797-1805. https://doi.org/10.1121/1.401660

Shaw, K. E., & Bortfeld, H. (2015). Sources of confusion in infant audiovisual speech perception research. *Frontiers in psychology*, *6*, 1844. https://doi.org/10.3389/fpsyg.2015.01844

Simpson, E. A., Varga, K., Frick, J. E., & Fragaszy, D. (2011). Infants experience perceptual narrowing for nonprimate faces. *Infancy*, *16*(3), 318–328. https://doi.org/10.1111/j.1532-7078.2010.00052.x

Siva, N., Stevens, E. B., Kuhl, P. K., & Meltzoff, A. N. (1995). A comparison between cerebral-palsied and normal adults in the perception of auditory-visual illusions. *The Journal of the Acoustical Society of America*, *98*(5_Supplement), 2983-2983. https://doi.org/10.1121/1.413907

Soderstrom, M., Casillas, M., Bergelson, E., Rosemberg, C., Alam, F., Warlaumont, A. S., & Bunce, J. (2021). Developing a cross-cultural annotation system and metacorpus for studying infants' real world language experience. *Collabra: Psychology*, *7*(1), 23445. https://doi.org/10.1525/collabra.23445

Stacey, J. E., Howard, C. J., Mitra, S., & Stacey, P. C. (2020). Audio-visual integration in noise: Influence of auditory and visual stimulus degradation on eye movements and perception of the McGurk effect. *Attention, Perception, & Psychophysics*, *82*(7), 3544–3557. https://doi.org/10.3758/s13414-020-02042-x

Sugden, N. A., Mohamed-Ali, M. I., & Moulson, M. C. (2014). I spy with my little eye: Typical, daily exposure to faces documented from a first-person infant perspective. *Developmental Psychobiology*, *56*(2), 249–261. https://doi.org/10.1002/dev.21183

Taitelbaum-Swead, R., & Fostick, L. (2016). Auditory and visual information in speech perception: A developmental perspective. *Clinical Linguistics & Phonetics*, *30*(7), 531–545. https://doi.org/10.3109/02699206.2016.1151938

Tiippana, K., Andersen, T. S., & Sams, M. (2004). Visual attention modulates audiovisual speech perception. *European Journal of Cognitive Psychology*, *16*(3), 457-472. https://doi.org/10.1080/09541440340000268

Tiippana, K., Ujiie, Y., Peromaa, T., & Takahashi, K. (2023). Investigation of cross-language and stimulus-dependent effects on the McGurk effect with Finnish and Japanese speakers and listeners. *Brain Sciences*, *13*(8), 1198. https://doi.org/10.3390/brainsci13081198

Ujiie, Y., & Takahashi, K. (2022). Own-race faces promote integrated audiovisual speech information. *Quarterly Journal of Experimental Psychology*, *75*(5), 924–935. https://doi.org/10.1177/17470218211044480

Ujiie, Y., Kanazawa, S., & Yamaguchi, M. K. (2020). The other-race-effect on audiovisual speech integration in infants: A NIRS study. *Frontiers in Psychology*, *11*, 971. https://doi.org/10.3389/fpsyg.2020.00971

Ujiie, Y., Kanazawa, S., & Yamaguchi, M. K. (2021). The other-race effect on the McGurk effect in infancy. *Attention, Perception, & Psychophysics*, *83*(7), 2924–2936. https://doi.org/10.3758/s13414-021-02342-w

UNICEF. (2019). *Family-friendly policies in rich countries: How Canada compares*. https://childcarecanada.org/documents/research-policy-practice/19/06/family-friendly-policies-rich-countries-how-canada-compares

Uttley, L., de Boisferon, A. H., Dupierrix, E., Lee, K., Quinn, P. C., Slater, A. M., & Pascalis, O. (2013). Six-month-old infants match other-race faces with a non-native language. *International Journal of Behavioral Development*, *37*(2), 84-89. https://doi.org/10.1177/0165025412467583

Valentine, T. (1991). A unified account of the effects of distinctiveness, inversion, and race in face recognition. *The Quarterly Journal of Experimental Psychology*, *43*(2), 161-204. https://doi.org/10.1080/14640749108400966

Vannasing, P., Dionne-Dostie, E., Tremblay, J., Paquette, N., Collignon, O., & Gallagher, A. (2024). Electrophysiological responses of audiovisual integration from infancy to adulthood. *Brain and Cognition*, *178*, 106180. https://doi.org/10.1016/j.bandc.2024.106180

Walker, A. S. (1982). Intermodal perception of expressive behaviors by human infants. *Journal of Experimental Child Psychology*, *33*(3), 514-535. https://doi.org/10.1016/0022-0965(82)90063-7

Walker-Andrews, A. S., Bahrick, L. E., Raglioni, S. S., & Diaz, I. (1991). Infants' bimodal perception of gender. *Ecological Psychology*, *3*(2), 55-75. https://doi.org/10.1207/s15326969eco0302_1

Wallace, M. T., & Stein, B. E. (1997). Development of multisensory neurons and multisensory integration in cat superior colliculus. *Journal of Neuroscience*, *17*(7), 2429-2444. https://doi.org/10.1523/JNEUROSCI.17-07-02429.1997

Wallace, M. T., Perrault, T. J., Hairston, W. D., & Stein, B. E. (2004). Visual experience is necessary for the development of multisensory integration. *Journal of Neuroscience*, *24*(43), 9580-9584. https://doi.org/10.1523/JNEUROSCI.2535-04.2004

Wallace, M. T., Woynaroski, T. G., & Stevenson, R. A. (2020). Multisensory integration as a window into orderly and disrupted cognition and communication. *Annual review of Psychology*, *71*(1), 193-219. https://doi.org/10.1146/annurev-psych-010419-051112

Weatherhead, D., & Werker, J. F. (2022). 20-month-olds selectively generalize newly learned word meanings based on cues to linguistic community membership. *Developmental Science*, *25*(4), e13234. https://doi.org/10.1111/desc.13234

Weikum, W. M., Vouloumanos, A., Navarra, J., Soto-Faraco, S., Sebastián-Gallés, N., & Werker, J. F. (2007). Visual Language Discrimination in Infancy. *Science*, *316*(5828), 1159–1159. https://doi.org/10.1126/science.1137686

Xiao, N. G., & Lee, K. (2018). iTemplate: A template-based eye movement data analysis approach. *Behavior Research Methods*, *50*(6), 2388–2398. https://doi.org/10.3758/s13428-018-1015-x

Xiao, N. G., Angeli, V., Fang, W., Manera, V., Liu, S., Castiello, U., ... & Simion, F. (2023). The discrimination of expressions in facial movements by infants: A study with point-light displays. *Journal of Experimental Child Psychology*, *232*, 105671. https://doi.org/10.1016/j.jecp.2023.105671

Yan, L., Liu, X., Hu, S., Liu, S., Krasotkina, A., & Xiao, G. N. (2025). Developmental Origins of Cultural Differences in Audiovisual Speech Integration: Evidence from Canadian and Chinese Infants. https://doi.org/10.31234/osf.io/xpuvj

Young, G. S., Merin, N., Rogers, S. J., & Ozonoff, S. (2009). Gaze behavior and affect at 6 months: Predicting clinical outcomes and language development in typically developing infants and infants at risk for autism. *Developmental Science*, *12*(5), 798–814. https://doi.org/10.1111/j.1467-7687.2009.00833.x

Yuan, J., Hu, X., Chen, J., Bodenhausen, G. V., & Fu, S. (2019). One of us? How facial and symbolic cues to own-versus other-race membership influence access to perceptual awareness. *Cognition*, *184*, 19-27. https://doi.org/10.1016/j.cognition.2018.12.003

Supplementary Materials

Syllable-specific Control Analysis.

Combinations of syllable types did not impact infants' looking behaviour

Although the four syllables used in this study —/ba/, /ga/, /pa/, and /ka/— are common across many world languages, including English, it is possible that infants may exhibit differential sensitivity to specific syllable pairings. Such sensitivity could introduce a potential confound, leading to differences in looking times that may be driven by syllable-specific effects rather than the infants' capacity to integrate auditory and visual speech information. To address this possibility, we conducted a syllable-specific control analysis, grouping trials based on the auditory syllable types into two categories: (1) /**BaGa**/ combinations, which included trials where the auditory syllable was either /ba/ or /ga/ (i.e., auditory /ba/-visual /ga/ and auditory /ga/-visual /ba/), and (2) /**PaKa**/ **combinations**, where the auditory syllable was either /pa/ or /ka/ (i.e., auditory /ka/-visual /pa/ and auditory /pa/-visual /ka/).

Data analyses revealed no significant contribution of syllable pairings to infants' looking behaviour. We first combined all infant participants' data from Experiments 1 and 2 and fitted a linear mixed-effects model with syllable combinations (/BaGa/ and /PaKa/) as a fixed effect and a random intercept for each participant to account for repeated-measures nature of the data, as each infant experienced both syllable combinations. The dependent variable was infants' looking time differences between the McGurk and non-McGurk trials, the proxy measure for audiovisual speech integration. Results indicated that the difference in looking times did not significantly vary as a function of syllable pairing ($\beta = –0.23$, $SE = 0.23$, $t(106) = –0.98$, $p = .332$), suggesting that the observed AV integration effect was not systematically influenced by the specific syllables used.

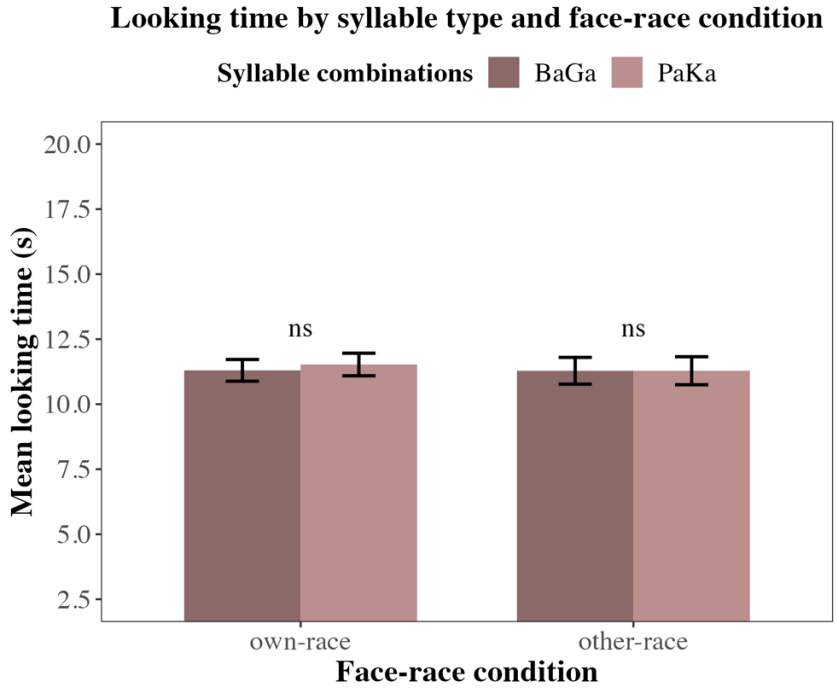
To further examine whether this null effect held consistently across experimental conditions, we included face-race condition (own-race vs. other-race) as an additional factor in a 2 × 2 linear mixed-effects model. This model included fixed effects for syllable pairings, face-race condition, and their interaction, as well as a random intercept for participant to account for the repeated-measures nature of the data. As expected, there was no significant main effect of syllable type (β = –0.30, SE = 0.23, t(104) = –1.32, p = .191). A significant main effect of face-race condition was found (β = –0.73, SE = 0.23, t(104) = –3.19, p = .002), such that overall looking preferences were attenuated in the other-race condition compared to the own-race condition, aligning with our finding that other-race faces impair audiovisual speech integration. However, this effect was independent of the ways in which syllables were paired in this study and does not implicate syllable type as a confounder, as there was no significant interaction between syllable type and face condition (β = –0.33, SE = 0.23, t(104) = –1.45, p = .152).

Finally, to further validate these findings, we conducted exploratory paired-sample *t*-tests within each face-race condition using both McGurk preference scores (i.e., looking time differences between both McGurk and non-McGurk trials) and total looking time in both trial types as dependent variables. First, we ran a set of paired-sample *t*-tests to examine whether the magnitude of infants' audiovisual speech integration—as indexed by McGurk preference scores—varied across the /BaGa/ and /PaKa/ syllable combinations. These tests revealed no significant effect of syllable combinations in either the own-race (t(32) = 0.09, p = .930) or other-race condition (t(20) = –1.49, p = .153), suggesting that infants' integration was not modulated by specific syllable pairings. Next, to test whether infants demonstrated greater attentional preference—as indexed by their total looking time—toward the two syllable combinations, we performed a second set of paired-sample *t*-tests using infants' mean total looking time as the

dependent variable, aggregated across both trial types. These results also revealed no significant difference in infants' overall looking to /BaGa/ versus /PaKa/ syllables in either the own-race ($t(32) = –0.67$, $p = .506$) or other-race ($t(20) = –0.00$, $p = .996$) condition (see Figure A).

Taken together, these results rule out the possibility that our findings from Experiments 1 and 2 could be attributed to a low-level attentional preference for a particular syllable category, strengthening our interpretation that infants' differential looking behaviour across McGurk and non-McGurk trials reflects genuine audiovisual speech integration.

Figure A. *Mean looking time for the /BaGa/ and /PaKa/ syllable combinations in the own-race (Experiment 1) and other-race (Experiment 2) conditions.*



Note. Error bars represent ±1 standard error of the mean. "ns" indicates non-significant differences ($p > .05$) based on paired-sample t-tests conducted separately within each face-race condition. No significant differences were found in infants' overall looking time between syllable combinations in either condition.

Gender Control Analysis

Gender does not make a difference on infants' looking preference.

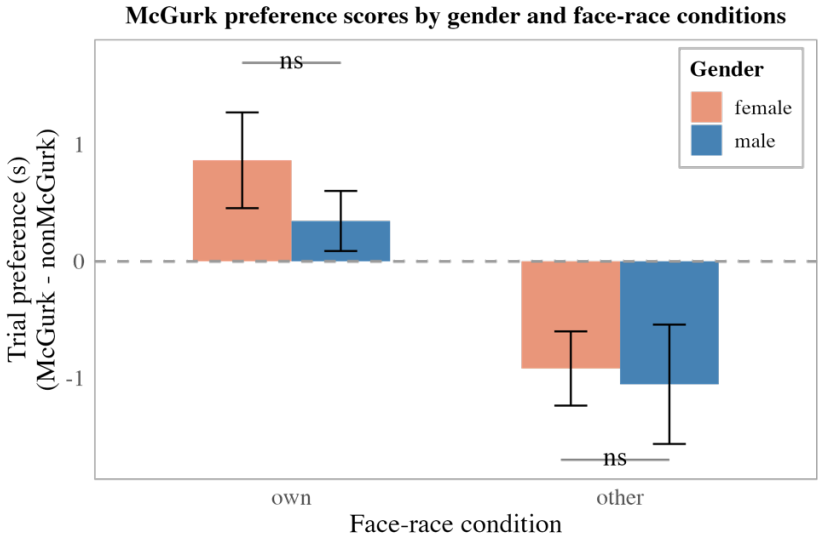
Previous literature has demonstrated gender differences in audiovisual speech integration in both infants (e.g., Desjardins & Werker, 2004) and adults (e.g., Irwin et al., 2006). To probe into whether such differences were present in the current study, we conducted series of statistical analyses using McGurk preference scores—calculated as the difference in looking time between McGurk and non-McGurk trials—as an index of infants' audiovisual speech integration.

We first combined data from own-race (Experiments 1) and other-race (Experiment 2) conditions. A multiple linear regression model was constructed with gender (male vs. female), face-race condition (own- vs. other-race), and their interaction as predictors. Type III sum of squares ANOVA revealed no statistically significant main effect of gender ($F(1, 50) = 0.76$, $p = .389$) nor gender × face-race condition interaction ($F(1, 50) = 0.26$, $p = .611$), suggesting that infants' gender did not influence the extent to which infants demonstrated audiovisual speech integration, and this null effect was consistent regardless of the race of the speaking faces. In contrast, a significant main effect of face-race condition ($F(1, 50) = 17.91$, $p < .001$) was observed, such that infants in the own-race condition exhibited greater McGurk preference scores than those in other-race condition, aligning with the previous finding in Experiments 1 and 2.

To more directly assess gender differences within each face-race condition, we conducted Welch's two-sample *t*-tests. Among infants in the own-race condition (Experiment 1), female infants ($M = 0.87$) did not significantly differ from male infants ($M = 0.35$) in their McGurk preference scores ($t(19.69) = 1.07$, $p = .296$). Similarly, no significant gender difference was observed in the other-race condition (Experiment 2) (female: $M = –0.92$; male: $M = –1.05$; $t(16.49) = 0.22$, $p = .825$). This absence of gender difference is visualized in Figure B below.

In sum, these findings indicate that while infants' audiovisual speech integration was modulated by the race of the speaking faces, there was no evidence of gender-based differences in the integration performance in either face-race condition. Thus, the observed other-race effect is independent of the gender of infant participants.

Figure B. *Trial looking preference by face-race condition and gender.*



Note. "ns" indicates a non-significant statistical result. Error bars represent ±1 standard error from the mean. The bars above the baseline represent longer looking time to the McGurk trials compared with non-McGurk trials, whereas the bars below the baseline reflect longer looking time to the non-McGurk than McGurk trials.

Non-White Infants' AV-Speech Integration

Non-white Canadian infants' lack of trial looking preference.

Looking data from the fourteen Canadian non-White infants who were excluded from Experiments 1 and 2 were grouped and analyzed together. Specifically, seven infants (Black, n = 1; biracial, n = 2 [Chinese and Pakistani; Black and White]; Hispanic, n = 1; Vietnamese, n = 1; Salvadoran, n = 1; Turkish, n = 1) were removed from Experiment 1 and thus viewed White

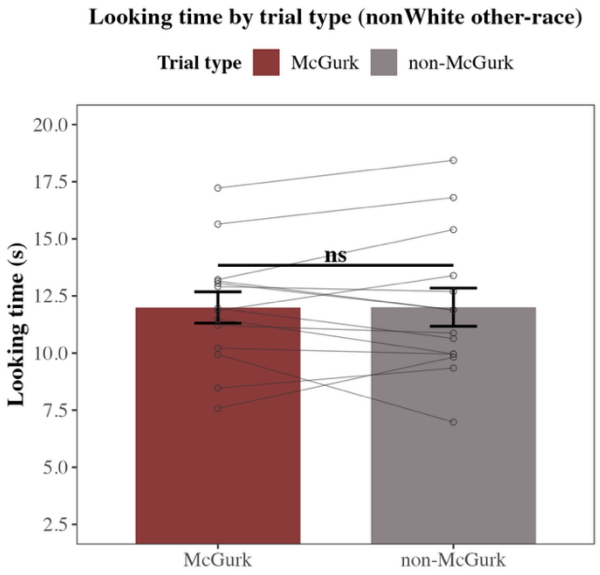
female faces, and eight infants (Arabic, $n = 1$; biracial, $n = 1$ [Black and White]; Filipino, $n = 3$; Indian, $n = 3$) were excluded from Experiment 2 and thus viewed East-Asian faces. Therefore, all infants were exposed to stimuli faces that fell into the other-race category, as the races of the stimulus faces—whether White or East-Asian—differed from that of their primary caregivers. One additional infant was excluded due to performance exceeding two standard deviations from the group mean.

Participants completed an average of 11 trials. As in all three experiments, we assessed whether these non-White Canadian infants demonstrated audiovisual speech integration when viewing other-race faces, by conducting a paired-sample *t*-test comparing their average looking times in the McGurk and non-McGurk trials. As shown in Figure C, infants looked equally at both types of trials ($M_{McGurk} = 11.99$ s, $SD_{McGurk} = 2.57$; $M_{non\text{-}McGurk} = 12.00$ s, $SD_{non\text{-}McGurk} = 3.13$; $t\,(12) = -1.42$, $p = .181$; Cohen's $d = -0.008$), indicating that their auditory perceptual outcomes were comparable across both types of trials. This suggested that their perception in both McGurk and non-McGurk trials was likely drive by the repetitive sequences of the physical auditory syllables (i.e., auditory /ba/, /pa/, /ga/, and/or /ka/). Therefore, these infants showed no evidence of audiovisual speech integration and did not appear to perceive an altered auditory percept in response to McGurk pairings.

We propose two possible explanations for the absence of evidence for audiovisual speech integration in this group of diverse non-White infants. First, the sample size may have been too small for a potentially significant preference to reach statistical significance. Second—and more plausibly—the other-race condition in this analysis was not as strictly controlled as in Experiment 2, making it less likely for a clean group-level preference to emerge. In other words, although all stimulus faces were technically other-race faces for these non-White infant

participants, the degree of deviation from each infant's most familiar face (i.e., their own-race face, defined here as the prototypical norm of their primary caregiver's racial category) varied too widely to be standardized (e.g., see Kelly et al., 2007; Valentine, 1991). For example, the perceptual distance between the stimulus East-Asian faces and the prototypical face norm for Filipino infants is likely smaller than that for Indian and Arabic infants. This variability in how other-race faces are encoded may have hindered the emergence of a measurable other-race effect on audiovisual speech integration. This consideration further supports our decision to exclude these infants from the defined own-race and other-race conditions in Experiments 1 and 2, respectively.

Figure C. *Mean looking times of non-White Canadian infants during McGurk and non-McGurk trials.*



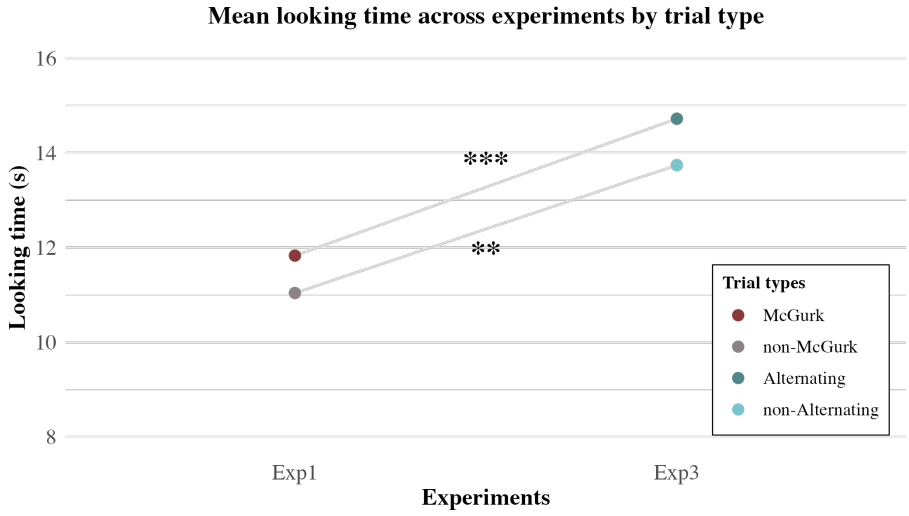
Note. "ns" indicates a non-significant statistical result. Error bars represent ±1 standard error from the mean. Each dot represents an individual participant's looking time in each trial type; lines connecting paired dots represent within-participant comparisons.

Infants' Preference for Congruent over Incongruent AV-pairings

Infants looked longer at congruent than mismatched audiovisual speech information.

In addition to trial-type preferences and visual attention allocation, we also explored whether infants' overall visual engagement—as indexed by looking time—differed between Experiment 1, which featured audiovisual incongruency, and Experiment 3, where auditory and visual syllables were always matched. A Welch two sample t-*t*est revealed that, across both Alternating and non-Alternating trials, infants in Experiment 3 exhibited significantly longer looking times than both trial types in Experiment 1. Specifically, the average overall fixation duration in Alternating trials of Experiment 3 (M = 14.72 s, SD = 3.23) was significantly greater than in McGurk trials of Experiment 1 (M = 11.72 s, SD = 2.99; t (42.86) = –3.40, $p < .001$, Cohen's d = –0.99). Similarly, the average fixation duration in non-Alternating trials (M = 13.74 s, SD = 3.02) exceeded that of non-McGurk trials (M = 11.04 s, SD = 3.25; t (49.60) = –3.19, p = .002, Cohen's d = –0.86). Figure D visualizes these two statistically significant increases in infants' overall fixation time. To ensure comparability, we only included the first eight of the total sixteen trials in Experiment 1, aligning with the number of trials with that of Experiment 3.

Figure D. *Mean overall looking times by both trial types across Experiments 1 and 3.*



Note. Each dot represents the averaged looking time for a given trial: McGurk (red) and non-McGurk (gray), as well as Alternating (darker blue) and non-Alternating (lighter blue). Colour coding of trial types is consistent with that in previous figures. Lines are drawn to connect comparable trial types across experiments for visual comparison. The asterisk(s) represent(s) the statistically significant differences in looking time between two trial types. No error bars are shown.

This pattern suggests that infants were more engaged overall when exposed to congruent audiovisual speech inputs, irrespective of the specific structures of the perceived auditory sequences. Whereas the mismatched auditory and visual syllables in Experiment 1 may have introduced perceptual ambiguity and placed greater demands on integrative processing, the fully matched audiovisual syllables in Experiment 3 may have facilitated bimodal speech processing and thus supported a more sustained attention. Although exploratory, this cross-experiment comparison points to audiovisual congruency as a potentially foundational factor in maintaining infants' attention to speech stimuli, extending our interpretations of audiovisual integration beyond trial-level preferences to include the broader influence of intersensory coherence.

PTLT-mouth by Trial Type and Face Condition Analysis

Infants looked more at the mouth during the non-McGurk trials.

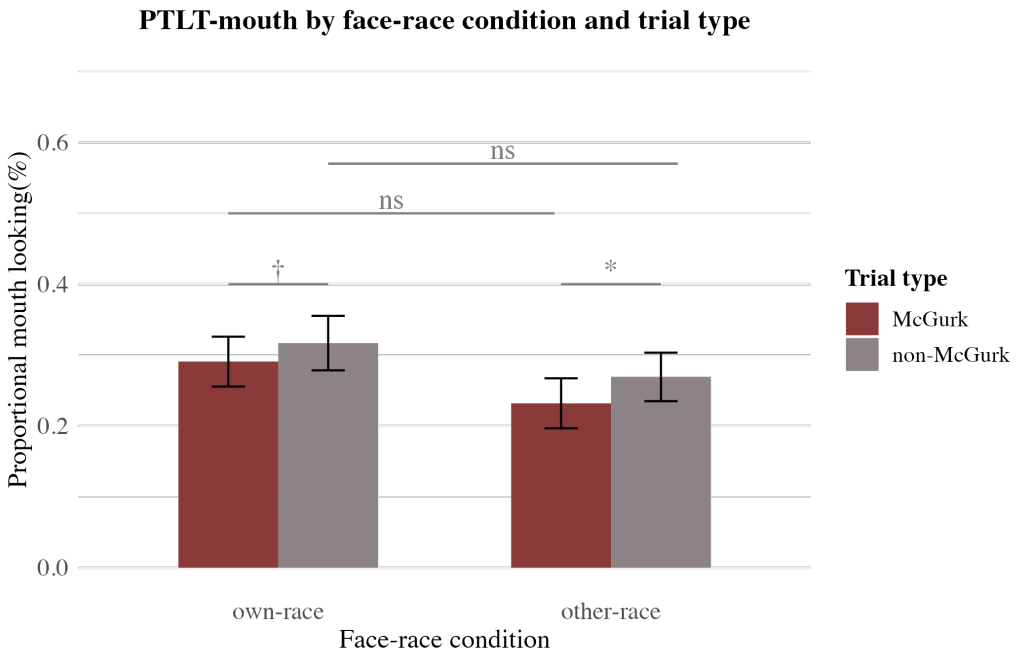
To justify our decision of collapsing data across trial types in the areas-of-interest (AOIs) analyses focusing on proportional looking to the mouth (PTLT-mouth), we examined whether infants' PTLT-mouth differed by trial types across the own- and other-race conditions using a set of linear mixed-effects models and follow-up Welch's two-sample *t*-tests.

A linear mixed-effect model using the full dataset combining Experiments 1 and 2 (i.e., collapsed across face-race condition) revealed a significant main effect of trial type on PTLT-mouth, with non-McGurk trials eliciting significantly greater mouth-looking than McGurk trials (b = –0.016, $t(52)$ = –3.16, p = .0026). Neither the main effect of face-race condition nor its interaction between trial type and face-race condition reached significance (both ps > .26), suggesting that the influence of trial type on PTLT-mouth was consistent across race conditions.

Separate analysis within each face-race condition offered further nuance. In the other-race condition, infants looked significantly more to the mouth during non-McGurk trials than McGurk trials (b = –0.0186, $t(20)$ = –2.79, p = .011). In the own-race condition, a similar but marginally significant trend was observed (b = –0.0131, $t(32)$ = –1.93, p = .062). Arguably, this consistent directionality may reflect a functional visual strategy independent from face-race. Specifically, although audiovisual inputs were conflicting in both trial types, the McGurk trials afforded fusible percepts where the auditory and visual syllables are *compatible* in mouth shapes; in contrast, the evidently *irreconcilable* auditory and visual syllables in the non-McGurk trials may have prompted infants to resolve by focusing more on the visual articulatory speech cues.

To assess between-group differences, we performed two Welch's two-sample t-tests to compare PTLT-mouth across face-race conditions within each trial type. Neither the McGurk trial comparison ($t(49.16)$ = 1.18, p = .243; Cohen's d = 0.31) nor the non-McGurk trial comparison ($t(51.28)$ = 0.93, p = .358; Cohen's d = 0.24) reached significance. This finding reinforces that trial type modulated PTLT-mouth but did not significantly differentiate across face-race conditions. Figure E visualizes these comparisons.

Figure E. *Proportional-of-total-looking to the mouth by trial type and face-race condition.*



Note. "ns" indicates a non-significant statistical result. Error bars represent ±1 standard error from the mean. The dagger "†" represents marginally significance ($.05 < p < 1.0$) and the asterisk represents statical significance (* $p < .05$).

Taken together, these results justify our analytical decision to collapse across trial types for the AOI analyses focused on PTLT-mouth. While trial type influenced PTLT-mouth—significantly in the other-race condition and marginally in the own-race condition—the effect was uniform in direction and did not interact with face-race condition. By collapsing across trial type, we aimed to provide a clear picture of the overall scanning patterns as a function of face-race familiarity.

Parental Questionnaire Data

Table A.

Averaged daily time (in hours) infants spent with primary caregivers, by experimental conditions.

	Mothers (hrs/day)	**Fathers (hrs/day)**	**Combined (hrs/day)**
own-race (Experiment 1)	21.77	14.04	33.47
other-race (Experiment 2)	20.61	9.74	29.30

Note. All values in this table were calculated by first averaging the reported total weekly hours that infants spent with their mothers, fathers, or both parents combined, and then dividing by 7 to obtain a daily estimate. Although all three values (i.e., with mothers, with fathers, and with both parents combined) were computed using the same method, the combined daily hours do not equal the sum of the mother and father averages. This discrepancy arises because the mother and father values were averaged separately across all available responses, and not all infants had data from both caregivers. As a result, the sum of the two individual caregiver averages is not mathematically identical to the average of combined weekly totals.