9-27-2021

Infant and Child Multisensory Attention Skills: Methods, Measures, and Language Outcomes

Elizabeth V. Edgar
Florida International University, eedga001@fiu.edu

FLORIDA INTERNATIONAL UNIVERSITY

Miami, Florida

INFANT AND CHILD MULTISENSORY ATTENTION SKILLS: METHODS, MEASURES, AND LANGUAGE OUTCOMES

A dissertation submitted in partial fulfillment of the

requirements for the degree of

DOCTOR OF PHILOSOPHY

in

PSYCHOLOGY

by

Elizabeth V. Edgar

2021

To: Dean Michael R. Heithaus
College of Arts, Sciences, and Education

This dissertation, written by Elizabeth Edgar, and entitled Infant and Child Multisensory Attention Skills: Methods, Measures, and Language Outcomes, having been approved in respect to style and intellectual content, is referred to you for judgment.

We have read this dissertation recommend that it be approved.

_____
Jacqueline Lynch

_____
Shannon Pruden

_____
Eliza Nelson

_____
Lorraine E. Bahrick, Major Professor

Date of Defense: September 27, 2021

The dissertation of Elizabeth Edgar is approved.

_____
Dean Michael R. Heithaus
College of Arts, Sciences, and Education

_____
Andrés G. Gil
Vice President for Research and Economic Development
and Dean of the University Graduate School

Florida International University, 2021

DEDICATION

For Karin and Doug, who saw my potential, encouraged

this journey, and made it all possible.

ACKNOWLEDGMENTS

This dissertation was made possible through the support and guidance of multiple individuals. I first wish to thank the members of my committee, Dr. Eliza Nelson, Dr. Shannon Pruden, and Dr. Jacqueline Lynch, for their help throughout the execution of this dissertation. I would like to thank my advisor Dr. Lorraine E. Bahrick for her mentorship and support throughout my doctoral studies. I would also like to thank the rest of my academic support system for their guidance and feedback: Dr. James Torrence Todd and Dr. Bret Eschman. Finally, I would like to thank Dr. Douglas Sellers and Dr. Karin Machluf for their early mentorship and for making the journey to my doctoral studies possible.

I also wish to thank Kaitlyn Testa, Kaityn Contino, Amanda Delgado, and Rossana Cardoso, who were my support system from my very first semester of graduate school and throughout the completion of my dissertation. I would like to thank the former and current members of Infant Development Lab for their support and dedication. I would like to thank Mark Fulton Jr., my soon-to-be husband, who moved across the country to join me in this journey, and who showed me unwavering support throughout my doctoral studies. Finally, I thank my mother and father for encouraging me to chase my dreams as a first-generation student, and supporting me while I did it.

My work was made possible through the UGS Dissertation Year Fellowship, which supported me throughout the last two semesters of my doctoral career. Thank you.

ABSTRACT OF THE DISSERTATION

INFANT AND CHILD MULTISENSORY ATTENTION SKILLS:

METHODS, MEASURES, AND LANGUAGE OUTCOMES

by

Elizabeth Edgar

Florida International University, 2021

Miami, Florida

Professor Lorraine E. Bahrick, Major Professor

Intersensory processing (e.g., matching sights and sounds based on audiovisual synchrony) is thought to be a foundation for more complex developmental outcomes including language. However, the body of research on intersensory processing is characterized by different measures, paradigms, and research questions, making comparisons across studies difficult. Therefore, Manuscript 1 provides a systematic review and synthesis of research on intersensory processing, integrating findings across multiple methods, along with recommendations for future research. This includes a call for a shift in the focus of intersensory processing research from that of assessing average performance of groups of infants, to one assessing individual differences in intersensory processing. Individual difference measures allow researchers to assess developmental trajectories and understand developmental pathways from basic skills to later outcomes. Bahrick and colleagues introduced the first two new individual difference measures of intersensory processing: The Multisensory Attention Assessment Protocol (MAAP) and The Intersensory Processing Efficiency Protocol (IPEP). My prior research using the MAAP has shown that accuracy of intersensory processing at 12 months of age predicted

18- and 24-month child language outcomes. Moreover, it predicted child language to a greater extent than well-established predictors, including parent language input and SES (Edgar et al., under review)! Manuscript 2 extends this research to examine both speed and accuracy of intersensory processing using the IPEP. A longitudinal sample of 103 infants were tested with the IPEP to assess relations between intersensory processing at 6 months of age and language outcomes at 18, 24, and 36 months, while controlling for traditional predictors, parent language input and SES. Results demonstrate that even at 6 months, intersensory processing predicts 18-, 24-, and 36-month child language skills, over and above the traditional predictors. This novel finding reveals the powerful role of intersensory processing in shaping language development and highlights the importance of incorporating individual differences in intersensory processing as a predictor in models of developmental pathways to language. In turn, these findings can inform interventions where intersensory processing can be used as an early screener for children at risk for language delays.

TABLE OF CONTENTS

LIST OF TABLES

TABLE PAGE

MANUSCRIPT 1

MANUSCRIPT 2

LIST OF FIGURES

FIGURE PAGE

MANUSCRIPT 1

MANUSCRIPT 2

I. GENERAL INTRODUCTION

The term “multisensory attention skills” was recently introduced by Bahrick and colleagues to encompass three basic skills: intersensory processing (matching sights and sounds based on audiovisual synchrony), sustained attention (maintaining attention to a stimulus, task, or event, often in the presence of competing stimulation), and speed of shifting/disengaging (shifting and/or disengaging visual attention to a new stimulus or location). Multisensory attention skills are fundamental for perceiving the sights, sounds, and/or tactile stimulation from a single event as unitary (Bahrick et al., 2020).

Intersensory processing is thus a critical foundation upon which more complex language, social, and cognitive skills can develop (Bahrick et al., 2020; Bahrick & Lickliter, 2012; Barutchu et al., 2019; Pons et al., 2019). However, the body of research on intersensory processing is characterized by different measures, paradigms, and research questions, making comparisons across studies difficult. Further, these studies are often designed for group-level analyses (e.g., assessing groups of infants at specific ages), making questions regarding the pathways from early developing, intersensory processing skills to later developing, complex skills difficult to address. Therefore, Manuscript 1 presents a broad, up-to-date review of findings in the area of intersensory processing, with a focus on behavioral methods used to assess audiovisual intersensory processing of faces and voices in infants and young children. This review will synthesize what is known across disparate methods and measures typically studied separately to provide a more comprehensive picture of the state of the field than is currently available. With this

foundation, we then call for a shift in the focus of research from one focusing on group differences to one assessing individual differences in intersensory processing skills.

Individual difference measures allow researchers to examine intersensory processing in individual infants and children and to relate those differences to individual differences in other skills such as language. Thus, developmental trajectories and pathways from early intersensory processing skills to later developmental outcomes can be explored. Recently, Bahrick and colleagues developed the first new individual difference measures appropriate for assessing intersensory processing in preverbal and/or nonverbal children or infants. The Multisensory Attention Assessment Protocol (MAAP; Bahrick, Todd, et al., 2018) assesses accuracy of intersensory processing along with sustained attention and speed of shifting. The Intersensory Processing Efficiency Protocol (IPEP; Bahrick, Soska, et al., 2018) is fine-grained measure of just intersensory processing speed and accuracy. Both protocols assess attention in the context of dynamic audiovisual events.

Research has only recently begun to examine the relation between early intersensory processing and later language skills using an individual differences approach. Specifically, my thesis research demonstrated that intersensory processing of faces and voices on the MAAP at 12 months of age predicts child quality and quantity of speech at 18 and 24 months, as well as expressive vocabulary size at 18 months, over and above the traditional predictors, parent language input (quality and quantity) and SES (Edgar et al., under review). Although intersensory processing on the MAAP has emerged as a predictor of child language, the IPEP, a more difficult and fine-grained measure of intersensory processing, has not yet been used to assess intersensory

processing and child language outcomes. The IPEP is capable of revealing smaller differences among infants than the MAAP. Small differences in early development can cascade to larger differences in later developing skills such as language.

Manuscript 2 of this dissertation examines intersensory processing at 6 months as a predictor of child language outcomes at 18, 24, and 36 months, over and above the traditional predictors, parent language (quality and quantity) and SES. In keeping with the call for a focus on individual difference measures, this paper builds directly on my previous research (Edgar et al., under review) by examining if language can be predicted from earlier intersensory processing skills (i.e., 6 instead of 12 months) and with a more fine-grained measure of intersensory processing (e.g., IPEP instead of MAAP). Similar to my thesis project, Manuscript 2 thus illustrates a shift in the focus of research to one of individual differences in intersensory processing as advocated by Manuscript 1. Using this approach, Manuscript 2 can reveal the ages at which intersensory processing best predicts child language outcomes, and can provide an understanding of the developmental processes and pathways leading to both typical and atypical development.

Together, the manuscripts included in this dissertation provide a comprehensive picture of the current state of knowledge of intersensory processing of faces and voices in infants and young children. This dissertation provides a broad, up-to-date review of the body of research on intersensory processing of faces and voices. It highlights the foundation provided by previous literature, the majority of which used a group difference approach, and calls for the use of an individual difference approach in future research (Manuscript 1). It then provides an empirical demonstration of how to use individual

difference measures to study developmental pathways from early intersensory processing skills to later developmental outcomes in the (Manuscript 2).

References

Bahrick, L. E., & Lickliter, R. (2012). The role of intersensory redundancy in early perceptual, cognitive, and social development. In A. Bremner, D. J. Lewkowicz, & C. Spence (Eds.), *Multisensory development* (pp. 183–205). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199586059.003.0008

Bahrick, L. E., Lickliter, R., & Todd, J. T. (2020). The development of multisensory attention skills: Individual differences, developmental outcomes, and applications. In J. J. Lockman & C. S. Tamis-LeMonda (Eds.), *The Cambridge Handbook of Infant Development* (pp. 303–338). Cambridge University Press.

Bahrick, L. E., Soska, K. C., & Todd, J. T. (2018). Assessing individual differences in the speed and accuracy of intersensory processing in young children: The Intersensory Processing Efficiency Protocol. *Developmental Psychology, 54*(12), 2226–2239. https://doi.org/10.1037/dev0000575

Bahrick, L. E., Todd, J. T., & Soska, K. C. (2018). The Multisensory Attention Assessment Protocol (MAAP): Characterizing individual differences in multisensory attention skills in infants and children and relations with language and cognition. *Developmental Psychology*, *54*(12), 2207–2225. https://doi.org/10.1037/dev0000594

Barutchu, A., Toohey, S., Shivdasani, M. N., Fifer, J. M., Crewther, S. G., Grayden, D. B., & Paolini, A. G. (2019). Multisensory perception and attention in school-age children. *Journal of Experimental Child Psychology*, *180*, 141–155. https://doi.org/10.1016/j.jecp.2018.11.021

Edgar, E. V., Todd, J. T., & Bahrick, L. E. (n.d.). *Intersensory matching of faces and voices in infancy predicts language outcomes in young children*. Manuscript under review.

Pons, F., Bosch, L., & Lewkowicz, D. J. (2019). Twelve-month-old infants' attention to the eyes of a talking face is associated with communication and social skills. *Infant Behavior and Development*, *54*, 80–84. https://doi.org/10.1016/j.infbeh.2018.12.003

II. MANUSCRIPT ONE

THE DEVELOPMENT OF INTERSENSORY PROCESSING OF FACES AND VOICES: A COMPREHENSIVE REVIEW AND ANALYSIS

Abstract

Intersensory processing (the detection of temporal synchrony, tempo, rhythm, duration across sensory modalities) emerges early in development and provides a foundation for later language, social, and cognitive outcomes. However, this body of research is characterized by different measures and paradigms which dictate the research questions and conclusions that can be drawn, making comparisons across studies difficult. This paper presents a comprehensive, up-to-date review of behavioral methods used to assess audiovisual intersensory processing of faces and voices in infants and children, synthesizing what is known across disparate methods and measures typically studied separately. We review six methods: intermodal preference, habituation, McGurk task, speech-in-noise, eye-tracking, and recent individual difference approaches. For each method, we review the research question(s) addressed, assumptions made, conclusions that can be drawn, research findings (including a table summarizing all studies) within each paradigm, and suggestions for future research directions. These foundational studies have predominantly used a group differences approach, which has generated a significant body of knowledge and set the stage for a new focus on individual differences in intersensory processing skills. This new approach allows researchers to assess predictive relations between intersensory processing skills and later outcomes and build models depicting how these basic skills cascade into later more complex language, social, and cognitive outcomes. The shift in focus can provide a greater understanding of the developmental processes and pathways leading to both typical and atypical development and provide a foundation for assessing infants at risk for later impairments in language, social, and cognitive functioning.

Introduction

The world of objects and events presents a dynamically changing flux of stimulation to all of our senses (e.g., changing sights, sounds, tactile impressions, etc.). Infants must learn to make sense of this dynamic multisensory stimulation from objects and events with no prior knowledge to guide them. They must learn to selectively attend to the dimensions of stimulation that optimize meaningful perception and action, while filtering out stimulation that is less relevant to their needs, goals, or interests at that moment in time (Bahrick et al., 2020). One feature of stimulation that guides attentional allocation to meaningful information in early infancy is intersensory redundancy (the synchronous co-occurrence of stimulation across two or more senses). Social events, including the faces and voices of people speaking, provide a rich source of intersensory redundancy. This recruits selective attention to amodal information (properties that are redundant across sensory modalities and not specific to a single sensory modality), including temporal synchrony (i.e., simultaneous changes in patterns of visual and acoustic stimulation, including auditory and visual onset, offset, duration, and common temporal patterning), rhythm, tempo, and intensity covariation. Amodal information thus organizes and directs selective attention, creating attentional salience hierarchies that allow infants to attend to global information prior to more specific detail (Bahrick, 2000, 2001; Bahrick & Lickliter, 2002).

The detection of redundant amodal information, or intersensory processing, has been found to be a cornerstone of early perceptual development (Bahrick & Lickliter, 2002; Lewkowicz, 2000; Lewkowicz & Lickliter, 1994), and proposed to provide a foundation for later social, cognitive, and language outcomes (Bahrick et al., 2020;

Bahrick & Lickliter, 2012). Infants selectively attend to amodal information in multimodal events, and this provides a basis for what is perceived, and in turn, what can be learned and remembered (Bahrick & Lickliter, 2014). Research assessing infant detection of amodal information has been conducted using different stimulus events, including both social events (e.g., faces and voices, people performing actions) and nonsocial events (objects impacting a surface). The body of research on intersensory processing is also characterized by different measures and paradigms which dictate the research questions and conclusions that can be drawn, making comparisons across studies difficult.

Here, we present a review of the behavioral methods used to assess intersensory processing of faces and voices in infants and young children. First, we provide a general overview of theory. Next, we discuss the various paradigms used to assess intersensory processing, including the intermodal preference method, habituation method, the McGurk task, the speech-in-noise task, eye-tracking, and new individual difference measures. We also created a table summarizing all of the studies and findings within each paradigm to serve as a guide for developmental scientists conducting multisensory research. Finally, we conclude with overall recommendations for future research directions.

Theoretical Background

Development of Intersensory Processing

Multisensory events such as the face and voice of a person speaking, provide information for a variety of amodal properties specified by common facial and vocal information (e.g., synchrony, rhythm, tempo, affect, prosody) and modality-specific

properties (e.g., pitch and timbre of the voice; configuration of facial features; color of skin and hair; (Bahrick et al., 2014; Bahrick et al., 2020). Events provide nested levels of amodal structure from global (temporal macrostructure, such as the overall temporal synchrony between the movements of the face and sounds of speech) to increasingly more specific levels (e.g., rhythm and tempo of facial movements and sounds of speech) to even more specific levels (e.g., common spectral information for specific phonemes specified by facial movements and speech sounds). The development of intersensory processing is thought to progress from the detection of global amodal relations to the detection of more specific amodal relations (Bahrick, 2000, 2001, 2010; E. J. Gibson, 1969).

The most global level of temporal information, temporal synchrony (usually between onsets and offsets of sights and sounds in object or speech events) plays a significant role in guiding and directing early perceptual development (Bahrick, 1988, 2002; Bahrick & Lickliter, 2000; Gogate et al., 2000; Gogate & Bahrick, 1998; Lewkowicz, 2003) and has received significant research focus (Bahrick, 1983, 1987; Hillairet de Boisferon et al., 2017; Hyde et al., 2011). It is thought of as the "glue that binds stimulation across the senses" (Bahrick & Lickliter, 2002; Bahrick & Pickens, 1994; Lewkowicz, 2000) and specifies the unity of an audiovisual event. Once infants can attend to unified multimodal events, the differentiation of specific, or nested, amodal relations can then proceed.

Infants detect global audiovisual synchrony in both social and nonsocial events in early development. Newborn infants detect temporal synchrony between mouth movements and sounds as early as 1 to 3 days of age (Lewkowicz et al., 2010), and show

evidence of learning arbitrary audiovisual relations on the basis of temporal synchrony (i.e., object labelling contingent on infant looking) at 2 days of age (Slater et al., 1999). Between three and seven months, infants detect temporal synchrony between the sights and sounds of faces and voices during speech (Dodd, 1979; Pickens et al., 1994) and between the sights and sounds of objects impacting a surface (Bahrick, 1983; Lewkowicz, 1992; Spelke, 1979, 1981).

The progression of intersensory perception in order of increasing specificity has been demonstrated in the domain of nonsocial (object) events. Research indicates that infants detect temporal synchrony between the sights and sounds of object impacts by 4 weeks of age as well as at older ages (Bahrick, 2001). The temporal synchrony between an object's audiovisual impact with a surface makes multiple levels of nested amodal temporal structure apparent. This includes information about different properties of the object, such as its' substance (elastic vs. rigid), composition (single object vs. compound/cluster of objects), number, or weight (Bahrick, 2004). In studies assessing infants' detection of both temporal synchrony and object composition in the same stimulus events (single and compound objects striking a surface), findings reveal that infants detect temporal synchrony earlier in development (by 3 to 4 weeks of age) than nested information for object composition, which emerges around 7 weeks of age (Bahrick, 2001). These findings indicate that within the first 2 months of life, infants progress from detecting global amodal properties between the sights and sounds of an object's impact, to detecting specific, nested amodal properties specifying the object's composition, information that is detectable within each synchronous impact. In contrast,

little research has focused on intersensory perception in order of increasing specificity for social (audiovisual speech) events.

In sum, young infants are adept perceivers of amodal information in multimodal events, readily detecting global audiovisual synchrony relations in both object and speech events in early development, and more specific amodal relations in object events somewhat later in development. However, multimodal events not only provide redundant amodal information, but also provide nonredundant modality-specific information (attributes available only to a specific sensory modality, such as pitch, timbre, pattern, or color). Under what conditions do infants perceive amodal information versus modality-specific information? What guides their attention and perceptual differentiation of different properties of multimodal events?

The Intersensory Redundancy Hypothesis

The Intersensory Redundancy Hypothesis (IRH; Bahrick & Lickliter, 2000, 2002) provides an explanation for how attention is allocated to amodal and modality-specific properties of events. The IRH is a theory that characterizes how the detection of amodal information, or intersensory processing, guides selective attention in early infancy, and how this process is coordinated with the detection of modality-specific information. Intersensory processing constrains perceptual learning so that infants detect and differentiate global properties first, followed by more specific details (Bahrick, 2000, 2001; Bahrick & Lickliter, 2002). As described earlier, the detection of properties of stimulation develops in order of increasing specificity, from global, amodal properties, to nested amodal properties, to more specific modality-specific information. This developmental progression fosters appropriate generalization of learning because

modality-specific details that vary across events can be perceived in the context of more global amodal properties that are less variable (Bahrick, 2001, 2010). For example, early detection of synchrony between faces and voices during speech allows infants to detect face-voice unity prior to detecting which specific voice belongs with a specific face.

According to the IRH, amodal information is more salient and detected first during multimodal (e.g., audiovisual) exploration than unimodal exploration (i.e., intersensory facilitation; see Bahrick & Lickliter, 2012). In multimodal stimulation, amodal information is highly salient and "tells" infants which properties of objects and events to attend to first and which to ignore. Amodal information guides attention to amodal properties of events at the cost of specific detail. For example, when presented with two visually superimposed videos (e.g., hands clapping and wooden sticks playing a toy xylophone), infants selectively attend to the video that is synchronized to its natural soundtrack, because it appears to "pop out" from the background of the other silent superimposed visual event (Bahrick et al., 1981). Infants also detect and learn to match objects and sounds on the basis of audiovisual information for object composition (e.g., single objects produce single impact sounds and compound object produce complex impact sounds) only if the objects are presented moving in synchrony with their impact sounds and not if they are presented out of synchrony (Bahrick, 1988). Further, ERP and heart rate measures have demonstrated amodal information, such as that available in synchronous faces and voices, is processed longer and more deeply than voices alone or the same faces and voices presented out of synchrony (Curtindale et al., 2019; Reynolds et al., 2014).

In contrast, modality-specific information is best detected during bouts of unimodal exploration (i.e., using just one sense, such as hearing a person speak from another room or observing them while silent). This occurs because attention is not captured by salient intersensory redundancy and is thus free to focus on modality-specific properties (e.g., unimodal facilitation; see Bahrick & Lickliter, 2012). Thus, modality-specific properties such as the specific acoustic qualities of the voice including timbre and pitch, or the configuration of facial features, are thought to be more salient and are detected first during unimodal exploration.

The early foundational research assessing principles of the IRH (e.g., intersensory and unimodal facilitation) focused primarily on infant detection of nonsocial (object) events. Infants show intersensory facilitation by the age of 3 to 5 months. They detect the amodal tempo and rhythm of a tapping toy hammer when the sights and sounds of the hammer are presented synchronously (i.e., intersensory facilitation), but not when the hammer event is presented without intersensory redundancy (e.g., unimodal visual, unimodal auditory, or asynchronous; Bahrick & Lickliter, 2000; Bahrick et al., 2002). In contrast, infants 3 and 5 months of age show unimodal facilitation of modality-specific information. They discriminate changes in the orientation of a tapping toy hammer when only the sight of the hammer is presented (i.e., unimodal visual), but not when the sights and sounds of the hammer were presented synchronously (Bahrick et al., 2006). The salience of temporal synchrony thus interfered with the detection of the modality-specific change in the orientation of the hammer.

More recently, research assessing principles of the IRH has focused on social events, including the faces and voices of people speaking. Social events are typically

more complex and variable than nonsocial events (Adolphs, 2001; Dawson et al., 2004) and provide an extraordinary amount of intersensory redundancy from rapidly changing coordinated patterns across face, voice, and gesture (Bahrick & Todd, 2012). Social events are highly salient and social interactions with caretakers guide and scaffold perceptual language, and social development, including social reciprocity, and mapping and learning new words (Bahrick, 2010). Research demonstrates that young infants show intersensory and unimodal facilitation of social events. For example, discrimination of affect (i.e., specified by a combination of amodal properties including intensity and tempo changes) is facilitated in bimodal audiovisual stimulation (i.e., intersensory facilitation), and impaired when it is conveyed in unimodal auditory and unimodal visual stimulation (i.e., without intersensory redundancy; Flom & Bahrick, 2007). In contrast, facial recognition, supported by detecting modality-specific properties, is facilitated when viewing silent faces of women speaking (i.e., unimodal facilitation) and impaired in the context of synchronous audiovisual speech in children 3.5- to 4-years-old (Bahrick et al., 2014).

The IRH also predicts change across development. As efficiency of processing, flexibility of attention, and perceptual differentiation progress across development, both amodal and modality-specific properties can be detected in multimodal, redundant and in unimodal, nonredundant stimulation. Research in the domains of nonsocial and social events support this developmental progression in infants (Bahrick & Lickliter, 2004; Flom & Bahrick, 2007, 2010). For example, at 4 months, infants can only detect affect in the context of synchronous audiovisual speech. At 5 months, they can detect affect in audiovisual synchronous speech as well as unimodal auditory speech, and by 7 months,

they can detect it under all conditions, in audiovisual synchronous speech, unimodal auditory speech, and unimodal visual speech (Flom & Bahrick, 2007). Intersensory facilitation and unimodal facilitation can also be seen in later development when tasks are difficult in relation to the expertise of the perceiver (Bahrick & Lickliter, 2004; Bahrick et al., 2010).

The Current State of Research on Intersensory Processing

Research on intersensory processing assessing infants' detection of dynamic, multimodal events abounds. Since the mid 1970s, a significant body of research has accrued demonstrating that infants have a wide range of intersensory processing skills in early development (for reviews see, Bahrick et al., 2020; Bahrick & Lickliter, 2012; Gogate & Hollich, 2010; Lewkowicz, 2000; Mason et al., 2019; Walker-Andrews, 1997). Further, it has generated important theory and conceptual development. This research has largely been informed by a group differences approach, in which groups of infants are tested and data are averaged across infants to characterize intersensory processing skills at specific ages (e.g., Bahrick, 2002; Bahrick et al., 1981; Patterson & Werker, 1999, 2003; Soken & Pick, 1992; Spelke et al., 1983). This research has generated a significant knowledge base about what intersensory skills infants demonstrate at different ages, as well as serving as an important foundation for the development of theory and indirect evidence indicating that intersensory processing skills provide a foundation for more complex developmental outcomes (Bahrick et al., 2020; Bahrick & Lickliter, 2012).

However, a group difference approach is limited in its ability to address questions about developmental processes and how foundational intersensory processing skills develop, are refined and cascade to other later developing competencies such as language

social and cognitive functioning. The research generated from this approach has provided a substantial knowledge base and strong foundation for addressing these questions. However, important questions addressing mechanisms of development and assessing developmental trajectories and pathways between basic intersensory skills and later developmental outcomes can only be addressed using an individual differences approach. Investigation of other skills including speech processing efficiency (Fernald et al., 2008) and visual attention and recognition memory (Rose et al., 2012) have shown significant advances because of the use of individual difference measures and longitudinal approaches. Using an individual difference approach, the intersensory processing skills of individual infants and children can be assessed relative to one another and these skills can then be related to their performance on later emerging skills such as language or cognitive tasks. These new questions require longitudinal approaches and individual difference measures that are fine-grained enough to assess intersensory processing skills of individual participants relative to one another.

To address this need, we have recently developed two new individual differences measures of intersensory processing skills appropriate for infants and children of all ages (Bahrick, Soska, et al., 2018; Bahrick, Todd, et al., 2018). They provide sufficiently fine-grained and reliable measures of intersensory processing for assessing developmental change in infant intersensory processing skills and relations with later outcomes such as language and social functioning. A number of studies have now been conducted using these measures (Bahrick, Todd, et al., 2018; Edgar et al., under review). In contrast with the indirect evidence afforded by group differences approaches, these studies have

demonstrated direct evidence of links between intersensory processing skills and later outcomes such as language development.

In sum, intersensory processing organizes and guides early perceptual development. Detection of amodal information creates attentional salience hierarchies that guide infant selective attention to information that is meaningful and relevant to their needs, goals, and actions. Decades of foundational research using primarily a group differences approach has generated a substantial body of knowledge about the intersensory processing skills of children at different ages. However, this body of research assesses intersensory processing using different measures and paradigms that each differ in terms of the research questions that can be asked and the conclusions that can be drawn, making findings difficult to integrate and synthesize across paradigms.

To address this challenge, we present a comprehensive review of the empirical studies to date on the topic of intersensory processing, with a focus on behavioral methods used to assess perception of audiovisual relations between stimulation from faces and voices in infants and young children. Included are the following methods: intermodal preference, habituation method, the McGurk task, the speech-in-noise task, eye-tracking, and the new individual differences approach. For each method, we review (a) the research question(s) addressed, (b) the assumptions made (c) conclusions that can be drawn, (d) the similarities and differences between methods, (e) research findings within each method including a table summarizing all the studies and findings within each paradigm, and (f) suggestions for future research directions. This paper will thus present a broad, up-to-date review of findings in the area of intersensory processing and synthesize what is known across disparate methods and measures typically studied

separately. This will provide a more comprehensive picture of the state of knowledge and research on intersensory processing of faces and voices in infants and children than is currently available. With this foundation, we then formulate recommendations for future research directions.

Paradigms

Intermodal Preference Method

Method and Research Questions

The intermodal preference method, also known as the intermodal matching method (Bahrick, 1983, 1988; Spelke, 1976), assesses a child's ability to detect a relation between a soundtrack and an appropriate visual event. It features two visual events side-by-side, along with a single soundtrack that is appropriate to only one of the events (e.g., it is synchronous and/or shares the same rhythm or tempo) and inappropriate to the other event (e.g., it is asynchronous and/or depicts a different rhythm or tempo). It requires no verbal skills, and places low cognitive demands on participants. The general research question addressed by the intermodal preference method is whether infants or children show audiovisual intersensory matching on the basis of amodal properties (e.g., audiovisual synchrony, common tempo, rhythm, intensity changes) or by attributes defined by combination of amodal properties (e.g., emotion, prosody, speaker's gender or age). It is assumed that infants detect the relation between the soundtrack and the visual event if the group of infants looks to the matching display over the mismatching display (or vice versa) significantly more than expected by chance (50%). One can thus conclude that infants detect a relation between the soundtrack and visual event based on the audiovisual amodal information that links them (e.g., rhythm, tempo, emotion, spectral

information, temporal synchrony, etc.). In the following section, we review selective research conducted using the intermodal preference method. We also include a table providing a comprehensive review summarizing findings of all studies to date assessing intersensory processing of faces and voices using the intermodal preference method in infants and young children (see Table 1).

Research Findings

Detection of Global Amodal Properties. A primary focus of empirical studies using the intermodal preference method has been to examine what amodal properties infants detect and the specific ages at which they are detected (see Table 1). A range of studies focus on infant detection of temporal synchrony and demonstrate sensitivity to temporal synchrony within the first few days of life. For example, at 1 to 3 days of age, neonates showed a significant proportion of total looking time to the synchronous mouth movements and sounds produced by nonhuman primate species (Lewkowicz et al., 2010). Neonates 2 days of age have also matched speech sounds with mouth/jaw movements during continuous infant-directed speech (two videos of identical woman speaking different sentences side-by-side) in point line displays (preserving motion of the mouth/jaw and head, but removing articulatory information) on the basis of temporal synchrony (Guellaï et al., 2016). An early study testing 2.5-month-olds found that infants spent a larger percentage of time attending to naturalistic audiovisual speech when the faces and voices were presented in synchrony than when they were presented out of synchrony (Dodd, 1979). Infants of 5 to 15 months of age were able to match the face and voice of a woman speaking pseudo-words in both natural and sine wave speech (preserving temporal information and the correlation between acoustic energy and visible

articulators, but degrading phonetic information) on the basis of temporal synchrony (Baart et al., 2013). Further, at 12 to 14 months (but not at 4 or 8 to 10 months), infants matched the sound-synchronous facial and vocal information presented in a non-native language (Lewkowicz et al., 2015). Thus, infants detect temporal synchrony during audiovisual speech in the first days of life (in nonhuman species and point-line displays of unfamiliar individuals), and continue to develop these skills across infancy, detecting it across a variety of event types (e.g., sine wave speech, point-light displays, pseudo words, continuous speech in the native and non-native languages).

Several studies have also investigated the detection of spectral information specifying vowel sounds in the first half year of life. These studies typically control for temporal synchrony by presenting videos of two women speaking in synchrony with one another but articulating different speech sounds. Using this method Kuhl and Meltzoff (1982) found that infants 4.5 to 5 months of age showed a significant proportion of total looking time to the lip movements matching the appropriate one of two vowel sounds. This was evident for natural speech, but not when the natural speech was replaced with a tone (i.e., removing spectral information), demonstrating matching on the basis of spectral information in the vowel sounds. These findings have been replicated and extended using male speakers (Patterson & Werker, 1999), and testing younger infants 2 to 3 months of age (Patterson & Werker, 2003). Thus, infants detect spectral information specifying vowel sounds in the presence of temporal synchrony by 2 to 3 months of age.

Detection of Attributes Defined by a Combination of Amodal Properties. A number of studies have demonstrated that infants can also detect amodal properties such as prosody, affect, gender, and age, that are specified by a combination of amodal

properties (see Table 1). For example, sadness is conveyed by lower intensity sounds and movements with slower tempo than happy affect. Similarly, prosody specifying approval is conveyed by higher intensity and more exaggerated rise-fall pitch contours than the prosody specifying prohibition (Bahrick et al., 2019). Gender and age are also specified by a combination of amodal properties. Males have a larger body (throat, face, chest, etc.) and a lower-pitched voice in a different formant frequency than females (Peterson & Barney, 1952), and children are characterized by smaller, rounder facial features, and voices of a higher pitch with a greater amplitude range than adults (Alley, 1981; Bahrick et al., 1998).

Only two studies have focused on the detection of prosody in audiovisual speech using the intermodal preference method. In addition to matching on the basis of temporal synchrony, the 2-day-old infants in Guellaï et al. (2016) matched sentences on the basis of audiovisual prosodic information (intonation, word stress, and phrasal rhythm) conveyed by the point-line displays of head motion and speech. Using the same method in an earlier study, infants 8 months of age likewise showed a significant proportion of total looking time to the head motions matching the vocal information in point-line displays on the basis of audiovisual prosodic information (Kitamura et al., 2014). These studies using the intermodal preference method indicate that infants match facial movement and vocal information in fluent speech on the basis of prosodic information even as neonates. However, further research is needed to determine which aspects of prosodic information infants detect at different ages.

A range of studies have focused on the detection of affective information across faces and voices in the first 7 months of life. By 7 (but not 5) months of age, infants show

a significant proportion of looking time to the display specifying common affect across the faces and voices of unfamiliar women for happy, sad, angry, and neutral emotions, when temporal synchrony is controlled (Soken & Pick, 1992; Walker-Andrews, 1986; Walker, 1982). They do so at 5 months when temporal synchrony is also available (Walker, 1982). Seven-month-olds can match on the basis of affect even when the lower half of the faces are occluded, showing only the eye region (preserving affect but removing the common rate of change across sounds and movements of the mouth; Walker-Andrews, 1986) and in point-line displays (preserving motion information, but removing body form; Soken & Pick, 1992). In contrast, infants showed no evidence of matching on the basis of affect when the faces were inverted (preserving visual pattern information, but making perception of configurational information more difficult; Walker, 1982), indicating that infants relied on configurational information for perceiving affect. Studies also demonstrate that infants as young as 3.5 to 4.5 months can detect affect in familiar faces (i.e., their mothers), but not in unfamiliar women when temporal synchrony is controlled (Kahana-Kalman & Walker-Andrews, 2001). Further, infants 5 months of age (but not 3.5 months) can match the facial and vocal emotional information of their peers (other infants) on the basis of affect (Vaillant-Molina et al., 2013). There have also been studies indicating that infants detect relations between static images of facial expressions and vocal information on the basis of affect (e.g., Flom et al., 2009; Flom & Whiteley, 2014). These studies, however, do not assess detection of amodal relations across faces and voices. Rather, they likely reflect the infant's prior knowledge of these relations, which could have developed on the basis of any number of processes (including learned associations, or detection of amodal relations). Thus, infants detect

common affect across faces and voices, including happy, sad, anger, and neutral expressions, during audiovisual speech consistently by 7 months of age, and do so earlier under specific conditions (e.g., for the faces of their mothers or other infants). They detect affect across a range of event types, including unfamiliar women, unfamiliar infants, and their mothers for naturalistic speech as well as point-light and static visual displays.

Several studies have also assessed the detection of amodal information for gender common across face and voice. Walker-Andrews et al. (1991) reported findings of two studies conducted in different labs. Both presented side-by-side displays of a male and female speaking in synchrony with each other side by side along with the voice of one or the other. Results converged and indicated that infants 6 to 6.5 months, but not 3 to 3.5 months, looked preferentially to the person matching the gender of the vocal information. In contrast, it has also been found that infants 6 months of age did not show a significant proportion of looking time to the gender matched facial and vocal information when a man and a woman sang a nursery rhyme in synchrony with each other using infant-directed speech (Hillairet de Boisferon et al., 2015), but did so when using adult-directed speech (Richoz et al., 2017). Similar discrepancies have been found at 9 months of age, for matching nursery rhymes sung by men and women, where infants matched on the basis of gender only for trials when the female was in sound (Hillairet de Boisferon et al., 2015), but matched for both genders in both adult- and infant-directed nursery rhymes in another study (Richoz et al., 2017). Gender matching has also been shown at various ages (5 to 24 months) for static images of males and females, indicating prior learning of relations between the sound of the voice and the visual appearance of the face (Lasky et

al., 1974; Poulin-Dubois et al., 1994, 1998). Infants can detect amodal information for gender in unfamiliar adults by 6 months. However, the conditions under which they detect gender (speaking vs. singing, infant- vs. adult-directed speech) at different ages remain unclear, likely due to differences across stimuli.

Studies of infant detection of age (e.g. child versus adult) based on amodal information common across faces and voices have also been conducted. One showed side-by-side videos of an adult and a child speaking in synchrony along with the voice appropriate to one (similar in design to the gender studies) and found that infants of both 4 and 7 months showed matching based on age. They matched when the faces were presented upright, but not inverted, indicating that facial configuration information is important for matching (Bahrick et al., 1998). An early study (Lasky et al., 1974) assessing matching based on age used voices along with static images and found that infants of 5 and 7 months showed matching but only for pictures of a woman paired with a boy and not for a man paired with a boy or for a woman paired with a man. Again, studies using static images assess prior learning and thus the basis for learning is not clear. Thus, one study demonstrates that infants can detect amodal information for age common across faces and voices by the age of 4 months.

Face-Voice Matching in Pre-Term Infants and Children Displaying Developmental Disabilities. Although the majority of studies using the intermodal preference paradigm have assessed intersensory perception in typically developing infants and children, a few have also focused on infants born pre-term and children displaying developmental disabilities. This research has examined whether infants and children show deficits in the detection of invariant temporal information compared to

typically developing infants and children (see Table 1). For example, when presented with side-by-side videos of two different women reciting a different song, one synchronized with the soundtrack, infants born pre-term showed no evidence of preferential looking to the sound-synchronized face at 3, 5, or 7 months of age. In contrast, infants born full-term showed matching on the basis of synchrony at 3 and 7 months of age (Pickens et al., 1994). Similarly, when presented with side-by side videos of the same woman speaking two different sentences in Japanese, one in synchrony with the soundtrack, Japanese pre-term infants at 6, 12, or 18 months of age, showed no preferential looking to the sound-synchronized event. In contrast, a group of infants born full-term showed matching at 6 and 18 months of age (Imafuku et al., 2019). Across the first year and a half of life, infants born pre-term show deficits in looking to sound-synchronous faces during continuous speech compared to infants born full-term.

A range of studies have examined intermodal matching in children with autism. Children with autism ranging from 2 to 7 years of age displayed a deficit in intermodal matching of faces and voices based on a variety of amodal properties including temporal synchrony and audiovisual information for affect, compared to typically developing children (Bebko et al., 2006; Kahana-Kalman & Goldman, 2007). Children with autism also displayed a deficit in intermodal matching of faces and voices when the display that was synchronous with the soundtrack was paired side-by-side with a display that had different levels of asynchrony between the face and voice (i.e., 0-s, 0.3-s, 0.6-s, 1-s; Righi et al., 2018). In contrast, children with autism show intermodal matching on par with that of typically developing infants under certain conditions. For example, they show preferential looking to the face of their mothers based on affect conveyed by her voice

(Kahana-Kalman & Goldman, 2007). Further, children with autism showed a significant proportion of total looking time to the temporally synchronous woman speaking while bouncing a doll in synchrony when the woman's face was obscured, but were unable to do so when her face was visible, indicating that children with autism detect synchrony but not in the context of a human face (Patten et al., 2016). From toddlerhood through early childhood, children with autism display deficits in face-voice matching of audiovisual speech compared to their typically developing peers. However, research using the intermodal preference method has not examined infants under 26 months of age (before autism is typically diagnosed), or in infants at low- versus high-risk for autism.

Predicting Developmental Outcomes. One recent study has also assessed the relation between face-voice matching in the intermodal preference method and concurrent language skills in young children with autism spectrum disorder (ASD) and language-matched typically developing (TD) children (see Table 1). Righi et al. (2018) assessed temporal synchrony perception in continuous speech and found that the proportion of total looking time to the face synchronized with vocal information was related to measures of receptive and expressive language in the TD children 3 years of age when the audiovisual asynchrony differed by 1-s (but not by 0-s, 0.3-s, or 0.6-s) from the synchronous display. This research suggests a direct link between temporal synchrony detection and langauge skills in young children. However, the intermodal preference method is not designed for examining individual differences, typically has low reliability, and thus presents a challenge for predicting outcomes.

Theoretical Constructs. The intermodal preference method has also been used to investigate theoretical constructs including perceptual narrowing (see Table 1).

Perceptual narrowing is a developmental process whereby infants show broad or unspecified perception of stimulus properties that gradually becomes more specific or attuned to their experience (Oakes & Rakison, 2020). Research examining perceptual narrowing with unimodal auditory stimuli shows that during the first few months of life, infants can discriminate phonetic contrasts in both their native and non-languages. But with increasing experience with their native language, infants display a decline in discrimination of non-native phonemes and improved discrimination of native phonemes (see Werker, 2018 for a review). Research on perceptual narrowing has also been conducted using audiovisual events. A study assessing perception of native versus non-native speech with temporal synchrony of speech in the two languages controlled, indicates an increase across age (5 to 10 months) in the proportion of total looking time to the matching facial and vocal information for native speech and a decrease for non-native speech (Shaw et al., 2015), consistent with predictions of perceptual narrowing. Using a modified intermodal preference method, with temporal synchrony controlled, a study demonstrated that infants of 4.5 months showed preferential looking to the face using both the native and non-native languages (Kubicek et al., 2014). In contrast, several studies found that later in development (at 6 and 10 to 14 months of age), infants preferentially looked to the face of the woman speaking that matched the vocal information in their native language, but not in the non-native language, demonstrating perceptual narrowing with age (Kubicek et al., 2014; Lewkowicz et al., 2015; Lewkowicz & Pons, 2013). Later in development (between 12 and 14 months), infants detected the relation between the face and sounds of the non-native language when temporal synchrony was available to differentiate the visual events (Kubicek et al., 2014;

Lewkowicz et al., 2015). Perceptual narrowing for the sounds of one's native language in audiovisual speech appears to occur across the second half of the first year and mirrors findings from the domain of unimodal auditory speech.

Perceptual narrowing has also been demonstrated for the perception of sounds produced by nonhuman primates. For example, at 4 and 6 months of age infants matched the facial movements and vocalization (coos versus grunts) of nonhuman primates on the basis of temporal synchrony, but no longer did so at 8 and 10 months of age (Lewkowicz & Ghazanfar, 2006). The authors argue that this is presumably because intersensory processing is initially broadly tuned and accommodates all forms of amodal information, but over time, intersensory processing becomes attuned to the amodal information that is most relevant to species-specific needs and ecology, in this case the face-voice synchrony of other humans. This body of literature suggests that perceptual narrowing may occur for intersensory matching of faces (from nonhuman primate species to human faces and from native to non-native languages) in the second half of the first year.

The Intermodal Preference Method: Summary & Future Directions

Table 1 displays a summary of all studies examining infant matching of faces and voices on the basis of amodal information using the intermodal preference method. This method was designed to be used for group-level analyses assessing which properties common across faces and voices during audiovisual speech groups of infants of specific ages can detect. A rich and varied body of research has demonstrated that infants detect a variety of amodal properties during the first half year of life, including temporal synchrony, spectral information for vowel sounds, prosody, affect, gender, and age of a speaker. Infants born pre-term and children with autism show poorer intersensory

matching of faces and voices than their typically developing peers. Further, in the domain of language, there is evidence of intersensory matching of faces and voices based on spectral information for phonemes in early infancy as well as perceptual narrowing to the native language across the first year for face-voice matching during continuous speech. In contrast with the group differences approach typically used in studies of the intermodal preference paradigm, one recent study has treated matching scores as individual difference variables and linked them with performance on tests of language in 3-year-old children with ASD and TD controls. This provides some direct evidence of links between early detection of synchrony between mouth movements and the voice and concurrent language skills.

This paradigm has generated an important knowledge base for the development of theory, including perceptual narrowing, and characterizing differences between typical and atypically developing children. However, there are a number of fruitful future directions for research using the intermodal preference method. First, given a limited number of studies, further research is needed to clarify the ages and conditions under which infants can detect prosody, gender, and age on the basis of amodal information. Second, to learn more about the development of intersensory processing skills in atypical development, research should assess pre-term infants and infants at high risk for autism across the first two years of life. However, early identification of children at risk for autism based on intersensory processing skills requires an individual differences approach (described in the "New Individual Difference Approaches" section). Finally, to advance theory about developmental processes, future research is needed to determine if the detection of amodal properties occurs in order of increasing specificity in the domain

of social events, paralleling the developmental progressions documented in the domain of nonsocial events.

The Habituation Method

Method and Research Questions

The habituation method (Horowitz, 1974; Horowitz et al., 1972) assesses an infant's ability to discriminate a change between the stimuli presented during the habituation phase and the stimuli presented during the test phase. Typically, this method uses an infant-control procedure such that the infant controls the amount of time the stimuli are presented based on their own looking behavior. This method features a series of habituation trials that begin when the infant fixates the screen and terminate when the infant looks away for a set amount of time, usually 1.5 or 2-seconds. Typically, each habituation trial is identical and trials are administered until the infant's visual attention to the screen decreases to the habituation criterion (often consisting of a 50% reduction in visual attention on two consecutive trials relative to the mean looking time across the first two trials; see Bahrick et al., 2019; Bahrick & Pickens, 1988; Flom & Bahrick, 2007; Lewkowicz, 2003). Unlike the other methods, the infant-control procedure allows each infant to control the overall amount of time they are exposed to the habituation stimuli, ensuring that they are sufficiently "bored" and ready to show visual recovery (an increase in looking relative to their own habituation level) if they detect a difference between the habituation stimuli and those presented during the test phase. Often, after habituation is reached, infants receive several post-habituation trials (identical to the habituation trials) in order to reduce the likelihood of chance habituation and visual recovery is assessed relative to looking during these trials. Typically, infants in the experimental condition

receive novel stimuli during the test trials and those in the control condition receive no change, and visual recovery is compared across the two groups. Alternatively, the novel and familiar trials are presented as a within-subjects variable (e.g., in alternation) during test trials.

Although both the intermodal preference and habituation methods are appropriate for nonverbal infants, the habituation method is less demanding in terms of perceptual processing. The intermodal preference method requires that infants actively compare two events and selectively attend to one of them over another (i.e., matching) on the basis of auditory and visual relations. In contrast, the habituation method presents just one display, and thus can provide a more sensitive index of perceptual skills than the intermodal preference method. Moreover, because it less demanding in terms of perceptual processing, it can often demonstrate intersensory processing skills at earlier ages.

An assumption underlying the habituation method is that visual recovery reflects the infant's discrimination of the change between the habituated stimulus and the novel test stimulus. In studies of intersensory processing, the habituation stimuli consist of audiovisual events (e.g., a person speaking in synchrony with a soundtrack) and the test trials depict a change in audiovisual relations, otherwise keeping elements of the stimuli identical to habituation (e.g., the same person speaking out of synchrony with the same soundtrack). If infants show visual recovery to the change, it can then be concluded that they detect the *relation* between the sights and sounds based on the amodal information that differs between habituation and test trials (in this case the temporal synchrony). In the following section, we review selective research findings assessing intersensory

processing of faces and voices in studies using the habituation method and include a table providing a comprehensive review summarizing studies using this paradigm (see Table 2).

Research Findings

Detection of Global Amodal Properties. A primary focus of research using the habituation method to assess intersensory processing of audiovisual speech has been to examine whether infants can detect a change in global amodal properties (e.g., temporal synchrony) and more specific amodal properties (e.g. tempo, rhythm) from habituation to test (see Table 2). A significant body of this research has been conducted using nonsocial events, such as objects impacting a surface, and demonstrate that early in development (e.g., 2 to 5 months) infants discriminate changes in rhythm and tempo in synchronous audiovisual events and only later do they discriminate these changes in asynchronous or unimodal visual events (Bahrick et al., 2002; Bahrick & Lickliter, 2000, 2004). Within the domain of social events, the majority of these studies have assessed the ages and conditions under which infants can detect changes in temporal synchrony. For example, infants 4, 6, and 8 months of age showed significant visual recovery to a change in temporal information (from synchrony to asynchrony) when the test trials depicted a novel woman speaking novel syllables out of synchrony, but only infants 6 and 8 months of age showed visual recovery to the same woman speaking novel syllables out of synchrony (Lewkowicz, 2000). These findings indicate that by 4 months, infants detect a change in audiovisual speech sounds and temporal synchrony, but by 6 months, they detect changes in the identity of the woman, speech sounds, and temporal synchrony. Further, in a study designed to assess the threshold for detecting audiovisual temporal

asynchrony, Lewkowicz (2010) first habituated infants with the face of a woman speaking a syllable in synchrony with its sound and tested infants with the same woman speaking the syllable with increasing degrees of asynchrony (366, 500, 666 ms). Infants 4 to 10 months of age detected a change for only the largest degree of asynchrony (0-ms vs. 666-ms). Then, infants were habituated with the face of the woman speaking the syllable 666-ms out of synchrony with its sound and tested with the same woman speaking the syllable with decreasing degrees of asynchrony (500, 366, 0 ms). Infants 4 to 10 months of age detected the largest change (666-ms vs. 0-ms) and a smaller change (666-ms vs. 366-ms). Using a similar design to assess the threshold for detection of audiovisual temporal synchrony in continuous native and non-native speech, Pons & Lewkowicz (2014) found that infants 8 months of age showed significant visual recovery to a change to the largest temporal asynchronies (0-ms vs. 666-ms and 0-ms vs. 500-ms), but not the smallest (0-ms vs. 366-ms) in continuous native and non-native speech. Together, these studies indicate the thresholds at which infants can detect changes in temporal synchrony across the first year of life in individual syllables and continuous speech.

Another study assessed infant discrimination of both rhythm and synchrony in audiovisual speech. Following habituation to a synchronous syllable spoken in a rhythmic pattern, infants 4, 6, 8, and 10 months of age showed significant visual recovery to the synchronous syllable presented in a novel rhythm. However, only infants 10 months of age discriminated this change when the temporal synchrony of the syllable was disrupted (Lewkowicz, 2003). These findings suggest that in early development, by 4 months, temporal synchrony is necessary for perceiving audiovisual rhythm information, whereas by 10 months it is no longer necessary.

Detection of Attributes Defined by a Combination of Amodal Properties.

Infant discrimination of attributes defined by a combination of amodal properties (e.g., emotion, age, gender, prosody) has also been assessed using the habituation method (see Table 2). Although a large portion of this research assesses infant discrimination of auditory-only speech streams or visual-only static images, a smaller body of research assesses detection of this information in audiovisual events and is reviewed here.

There have been a number of studies focused on infant discrimination of affect. This body of research has demonstrated that infants 4 to 7 months of age showed significant visual recovery to changes in dynamic positive and negative affective expressions during synchronous audiovisual speech of unfamiliar women for emotions including sad, happy, and angry (Caron et al., 1988; Flom & Bahrick, 2007; Walker-Andrews & Grolnick, 1983). Flom and Bahrick (2007) tested the intersensory facilitation principle of the IRH in infants 3, 4, 5, and 7 months of age in the domain of affect perception. By 4 months, infants discriminated changes in affect when presented with audiovisual synchronous stimulation but not the other conditions, indicating intersensory facilitation of affect information. Across development, discrimination of affect extended to unimodal auditory stimulation (5 months), and both unimodal auditory and visual stimulation (7 months) indicating intersensory facilitation of affect information in early but not later development. Flom et al. (2018) conducted a study assessing how the degree of familiarization time affects discrimination of affect. At 3 months of age, infants did not consistently discriminate changes in affect with the typical 50% habituation criterion (consistent with findings of Flom & Bahrick, 2007), but when given a longer familiarization time provided by a 70% habituation criterion, they did show

discrimination. It has also been demonstrated that 5-month-old infants showed significant visual recovery to a change in vocal affect when accompanied by a static image of a face (but not a checkerboard; Walker-Andrews & Lennon, 1991), indicating that in early development, infants require the context of a face to discriminate changes in vocal affect. In early development, infants detect changes in affect in synchronous audiovisual events but not unimodal visual or asynchronous events, demonstrating intersensory facilitation of affect. Between 4 and 7 months of age, infants detect changes in the affect of unfamiliar women, including happy, sad, and angry expressions, during continuous and synchronous audiovisual speech. Given an extended habituation time, they can do so as early as 3 months of age.

Only two habituation studies have focused on the detection of prosody in audiovisual speech. The first was designed to test the intersensory facilitation principle of the IRH (Bahrick et al., 2019). It demonstrated that infants 4 months of age showed significant visual recovery to a change in prosody (approval vs. prohibition) when presented with audiovisual synchronous, but not unimodal auditory or audiovisual asynchronous speech. Thus, in infants 4 months of age, discrimination of prosody was facilitated by the presence of intersensory redundancy (Bahrick et al., 2019). The second study focused on the detection of amodal information specifying language membership (e.g., differences in stress, intonation pattern and rhythm; Bahrick & Pickens, 1988). Infants 5 months of age were habituated to a woman speaking one of two passages in either English or Spanish, and tested with a novel passage presented in a novel language, a novel passage presented in the habituated language, or no change. They showed significant visual recovery to a novel passage spoken in a novel language (English vs.

Spanish), but not a novel passage spoken in the familiar, habituated language. Thus, infants can detect changes in prosody from prohibition to approval and vice versa in audiovisual speech by 4 months of age, and can detect changes in audiovisual speech on the basis of language membership at 5 months of age.

Research has also assessed the detection of arbitrary audiovisual relations. Bahrick et al. (2005) assessed the detection of arbitrary face-voice relations; the relation between the unique voice of a person and their face. Infants were habituated to two separate face-voice pairings in alternating trials, and tested by switching the face-voice pairings. Infants 4 and 6, but not 2, months of age detected arbitrary face-voice relations in unfamiliar men and women. They showed significant visual recovery to a change in the audiovisual face-voice pairings (i.e., face A synchronized with voice B and vice versa). In another study, Gogate and Bahrick (1998) assessed detection of arbitrary syllable-object relations. After habituation to audiovisual vowel-object pairs, 7-month-old infants learned arbitrary syllable-object relations when they were habituated to the syllables spoken in synchrony with a moving object (using a naming and showing gesture), but not when the vocalizations were presented out of synchrony with the moving object, or when they were paired with a static object (Gogate & Bahrick, 1998). Further, infants remembered the arbitrary syllable-object relation four days later, but only when habituated with temporal synchrony uniting the vocalizations with the moving object (Gogate & Bahrick, 2001). Thus, infants discriminate arbitrary intermodal relations between the appearance of a face and the particular sound of a voice by 4 months of age. Further, intersensory redundancy provides a basis for learning and remembering arbitrary vowel-object relations in infants 7 months of age. An important

direction for future research will be to determine the ages and conditions under which infants detect the different types of relations, global synchrony in audiovisual speech, nested amodal information for affect and prosody, and more specific arbitrary audiovisual relations such as specific face-voice relations and whether developmental progressions in the social domain parallel those found for nonsocial events.

The Habituation Method: Summary & Future Directions

Table 2 displays a summary of all studies assessing intersensory processing of audiovisual speech events using the habituation method. Research using audiovisual events has demonstrated that across the first year, infants discriminate changes in a variety of amodal properties common to faces and voices during audiovisual speech, including temporal synchrony, rhythm, affect, prosody, arbitrary face-voice relations, and language membership. Further, research has demonstrated that intersensory redundancy facilitates word-mapping and memory for syllable-object relations. As reviewed earlier, the habituation method assesses infants' discrimination of any change in stimulation from habituation to test stimuli and is thus less demanding in terms of perceptual processing than the intermodal preference method. The intermodal preference method is more demanding in that it requires active comparison between two events and selectively attending to one event over another (i.e. matching) on the basis of audible and visual relations. As such, the habituation method can be a more sensitive index of perceptual skills (e.g. detect more subtle changes in rhythm or affect) than the intermodal preference method, and in some cases demonstrates skills at earlier ages.

Similar to the intermodal preference method, the body of research conducted using the habituation method has expanded our knowledge base about the intersensory

processing skills in groups of infants at specific ages and has laid a foundation for the development of theory. In the domain of object (nonsocial) events, the habituation method demonstrates that infants discriminate amodal information in order of increasing specificity, and the few studies in the domain of face-voice (social) events aligns with this interpretation. Further, several studies establish indirect links between the detection of temporal synchrony and language skills, including the importance of temporal synchrony for learning speech sound-object relations in early development, and for distinguishing between the speech of two languages.

There are a number of future directions for research using the habituation method. First, future research characterizing infant discrimination of audiovisual speech on the basis of global amodal relations (temporal synchrony), more specific amodal relations (affect, prosody), modality-specific arbitrary relations (specific face-voice relations, speech sound-object relations) is needed. Habituation studies in the domain of nonsocial events (not included in this review) demonstrate that infants discriminate amodal information in order of increasing specificity (Bahrick, 1983, 1987, 2001, 2004). Although the existing research focused on social events aligns with this interpretation, future research should directly test this developmental progression within the domain of audiovisual speech events. Second, studies are needed assessing how typically and atypically developing infants and children differ in their discrimination of global and nested amodal properties. Future research should examine how these groups differ in early infancy. Finally, future research should further examine intersensory processing as a foundation for learning and memory of arbitrary word-object relations or specific face-voice relations in groups of infants at a wider range of ages.

McGurk Task

Method & Research Questions

The McGurk task (McGurk & MacDonald, 1976) addresses a participant's ability to integrate mismatching auditory and visual information into a unified percept. It depicts the face of a person speaking an auditory syllable (e.g., "ba") in synchrony with an incongruent visual syllable (e.g., "ga") to produce the percept of a third syllable that is different from both the auditory and visual syllables (e.g., "da" or "tha"). The perception of the third syllable is called the McGurk effect or illusion and is considered evidence of audiovisual integration (Macdonald & McGurk, 1978). It occurs because the auditory and visual stimuli differ in terms of the place of articulation. The auditory labials and visual non-labials produce a fused response of the emergent consonant (e.g., "d" or "th"). However, reversing the pairing of the auditory and visual syllables (e.g., auditory "ga" with an incongruent visual "ba") produces a combined percept (e.g., "bga") due to a heightened influence of the visual information. The McGurk task demonstrates the importance of the visual modality for speech perception. The general question assessed by the McGurk task is whether auditory and visual information are integrated when perceiving speech. In contrast with the intermodal preference and habituation methods which address a variety of research questions about intersensory processing depending on the stimulus contrasts and conditions presented, the McGurk task is designed to address a narrower research question regarding integrating or fusing auditory and visually presented syllables.

Both the intermodal preference method and habituation method have been used to assess whether infants perceive the McGurk effect in a manner similar to adults. For

older children and adults however, the McGurk task typically involves presenting a single display for the duration of a trial, unlike habituation. The response format also differs, as the intermodal preference and habituation methods do not require verbal responses, but instead assess looking behavior. In contrast, in older children and adults, the response format of the McGurk task requires a verbal, written, or button press response. It may also allow an open choice response (e.g., participants respond with any syllable they perceive) or a forced choice response (e.g., participants choose from a specific set of options).

Assumptions and conclusions made from the McGurk task differ somewhat depending on the methods used for testing infants or older children and adults. It is assumed that infants perceive the McGurk illusion if they look to the congruent and (e.g., auditory "ba" with visual "ba") and incongruent (e.g., auditory "ba" with visual "ga") displays for a similar amount of time (intermodal preference method) or if they do not increase look duration to the incongruent McGurk trials after habituation to congruent (e.g., auditory "ba" with visual "ba") trials (habituation method). It is assumed that older children and adults perceive the McGurk illusion if their forced choice button press response or verbal response indicates the emergent consonant, evidence for a unified percept. In the following section, we review selective research findings demonstrating evidence of the McGurk illusion as assessed by the McGurk task. We also include a comprehensive review summarizing all studies assessing intersensory processing of faces and voices using the McGurk task in infants and young children in Table 3.

Research Findings

Infant Perception of the McGurk Effect. A number of studies have examined whether infants perceive the McGurk effect (see Table 3) and most of these studies have used the habituation method. One study demonstrated that infants 5 months of age habituated to a congruent audiovisual syllable (e.g., auditory "va" with visual "va") showed significant visual recovery to the incongruent audiovisual syllable that elicits the McGurk effect (e.g., auditory "ba" with visual "va"), but not to an incongruent audiovisual syllable that does not elicit the McGurk effect (e.g., auditory "va" with visual "ba"; Rosenblum et al., 1997). These findings indicate detection of the McGurk/fusion effect (percept of a third syllable) rather than detection of any incongruency. Using a similar method, it has also been found that at 4 months of age, female, but not male, infants perceived the McGurk effect. However, after habituation to the incongruent audiovisual syllable that elicits the McGurk effect (e.g., auditory "bi" with visual "vi"), male infants showed significant visual recovery when tested with audiovisual "vi", suggesting that they perceived the McGurk effect in habituation (Desjardins & Werker, 2004). Further evidence of detecting the McGurk effect was found in infants 4 to 4.5 months of age (Burnham & Dodd, 1996, 2004). Infants habituated to a woman speaking the incongruent audiovisual syllable that elicits the McGurk effect showed a familiarity preference (indexed by visual fixation to a static image of the woman) when tested with the auditory-only fusion syllable (e.g., "da/tha"), whereas infants habituated to the congruent audiovisual syllable showed no familiarity preference for any of the syllables ("ba", "tha", "da"). Finally, when presented with an incongruent audiovisual syllable that elicits the McGurk effect (e.g., auditory "ba" with visual "va"), infants 5 to 5.5 months of

age showed no activation over the frontal and temporal areas in contrast with infants who were presented an incongruent audiovisual syllable that does not elicit the McGurk effect (e.g., auditory "va" with visual "ba"). Thus, infants perceived the mismatch between the auditory and visual incongruent syllables that could not be fused, but did not perceive a mismatch for syllables that could be fused (Kushnerenko et al., 2008). Infants perceive the McGurk effect in the first half year of life, indicating that, like adults, they are able to integrate auditory and visual speech information. This is supported by several different methods (habituation, ERP).

Child Perception of the McGurk Effect. A number of studies have examined the McGurk effect in young children (see Table 3). These studies assess McGurk perception using methodology similar to that of adults. Typically, a trial consists of a video of a person producing a syllable along with a synchronous soundtrack of a syllable. Different audible and visual combinations are presented (e.g., congruent audiovisual syllable, incongruent audiovisual syllable that elicits McGurk effect, incongruent audiovisual syllable that does not elicit McGurk effect). Participants then respond with the syllable they perceived, although response types (verbal vs. button-press) and formats (open choice vs. forced-choice) differ. Using this method, children 3 to 12 years of age show evidence of perceiving the McGurk effect (Dupont et al., 2005; Hirst et al., 2018; Massaro, 1984; McGurk & MacDonald, 1976; Sekiyama & Burnham, 2008; Tremblay et al., 2007). In children 6 to 12 years of age, FMRI measures showed greater activity in the left superior temporal sulcus (a critical region indicated in multisensory processing of audiovisual speech) for children who perceived the McGurk effect compared to children

who did not (Nath et al., 2011). Similar to infants, children perceive the McGurk effect, and this is supported by evidence from different methods (FMRI).

A range of studies has also compared child perception of the McGurk effect to that of adults. Children ranging from 3 to 11 years of age showed a significantly reduced ability to perceive the McGurk effect compared with adolescents and adults (Dupont et al., 2005; Hirst et al., 2018; Massaro, 1984; Tremblay et al., 2007). Children show reduced perception of the McGurk effect than adolescents and adults because they are less influenced by the visual stimulation (Dupont et al., 2005; Hirst et al., 2018; Massaro, 1984; McGurk & MacDonald, 1976; Sekiyama & Burnham, 2008), and more influenced by the auditory stimulation (Dupont et al., 2005; Hirst et al., 2018; Massaro, 1984). For example, when auditory noise (e.g., white noise played over soundtrack) or visual noise (e.g., blurring the video) was added to the presentation of the McGurk stimuli, children 3 to 6 years of age required a higher threshold of auditory noise to induce the McGurk effect and significantly less visual noise to eliminate the McGurk effect than adolescents and adults did (Hirst et al., 2018). Although children perceive the McGurk effect across the first decade, they show a reduced effect compared with adults, likely due to differential influence of the auditory and visual modalities which changes across age.

The McGurk effect has also been examined in other languages and compared to English-speaking participants (see Table 3). Although most of this research has been conducted with adults, Sekiyama & Burnham (2008) examined perception of the McGurk effect in Japanese- and English-speaking children. Children 6 years of age showed a relatively weak McGurk effect compared to adults, but with no significant difference between Japanese- and English-speaking children. By 8 years of age, perception of the

McGurk effect increased for the English-speaking children, but not for the Japanese-speaking children, likely due to differences in the speed of unimodal auditory processing. Japanese-speaking children had greater auditory-only speech perception than the English-speaking children. Thus, it appears that language experience has an impact on perception of the McGurk effect. Future research should examine child perception of the McGurk effect in languages more similar to English than Japanese, such as Dutch, German, or romance languages (Spanish, French, Italian) to determine if there are differences in processing of audiovisual speech.

Perception of the McGurk Effect in Infants at Risk for Autism and Children Displaying Developmental Disabilities. Research has also focused on perception of the McGurk effect in infants at high risk for autism and children displaying developmental disabilities. This body of research has assessed whether these children show a deficit in perception of the McGurk effect compared to TD children (see Table 3). One study has examined infants at high risk for autism and compared them to infants at low risk for autism. Using the intermodal preference method, infants 9 months of age at low risk for autism showed evidence of perceiving the McGurk effect. They looked to the congruent (e.g., visual "ga" with auditory "ga") and incongruent, fusible (e.g., auditory "ga" with visual "ba") displays for equal amounts of time, suggesting that they fused both auditory and visual syllable pairs. In contrast, infants at high risk for autism looked significantly longer to the incongruent display (e.g., auditory "ga" with visual "ba"), suggesting that they perceived the audiovisual mismatch and were not able to fuse the sounds to produce the McGurk effect (Guiraud et al., 2012). In the second half of the first year, infants at

high risk for autism display difficulty in integrating audiovisual speech compared to infants at low risk for autism.

A number of studies have assessed perception of the McGurk effect in children with autism. Children with autism ranging from 5 to 18 years of age showed a reduced perception of the McGurk effect compared with their typically developing peers (Irwin et al., 2011; Mongillo et al., 2008; Stevenson et al., 2014). This may be due to differences in unimodal processing. For example, children with autism perform significantly less accurately on unimodal visual trials (i.e., speech- or lip-reading) than typically developing children. After controlling for unimodal visual differences, differences in McGurk perception were no longer evident for children with autism versus TD children (Iarocci et al., 2010; Williams et al., 2004). At 7 years of age, children with autism showed reduced perception of the McGurk effect compared with their typically developing peers, but by 16 years of age, there were no significant differences in perception of the McGurk effect (Taylor et al., 2010). These findings suggest that perception of the McGurk effect (i.e., audiovisual integration) improves across age for children with autism. In contrast, it has been found that 13- to 18-year-olds with autism did not significantly differ from 6- to 12-year-old TD children in the rate at which they perceived the McGurk effect, and gave significantly less McGurk responses than 13- to 18-year-old TD children (Stevenson et al., 2014). These findings suggest that children with autism show slower developmental growth in intersensory processing of the McGurk effect across age. Children with autism show reduced perception of the McGurk effect compared with TD children, and differences appear to be influenced by poorer processing of unimodal visual speech. However, children with autism show

developmental growth in audiovisual integration of the McGurk effect, but at a slower rate than that of TD children.

Some studies have also focused on perception of the McGurk effect in children with language delays or disabilities. Children 3 to 5 years of age with atypical speech development (scored more than one standard deviation below the mean on a standardized speech assessment) did not differ significantly from TD children in the perception of the McGurk effect. However, children with delayed phonological development perceived the McGurk effect significantly more often than children with a diagnosed phonological disorder (Dodd et al., 2008). In contrast, 4- to 7-year-old children with selective language impairment (SLI; characterized by basic perceptual deficits in rapid auditory transitions) perceived the McGurk effect significantly less often than TD children, but across both groups of children greater language proficiency was related to greater perception of the McGurk effect (Norrix et al., 2007). Perception of the McGurk effect in children with language delays/disabilities varies as a function of the specific delay/disability and the degree of language proficiency they exhibit.

Perception of McGurk Effect in Relation to Developmental Outcomes. A few studies have examined how the perception of the McGurk effect relates to later outcomes (see Table 3). Boliek et al. (2010) assessed the relation between perception of the McGurk effect and math and reading achievement scores in children with a learning disorder (LD) and age-matched TD children 6 to 9 years of age. The children with LD who perceived the McGurk effect less frequently had lower reading and math achievement scores. In contrast, the TD children who perceived the McGurk effect more frequently had greater reading and math achievement scores. Further, in children 7 to 13

years of age, greater frequency of perceiving the McGurk effect was related to greater scores on an audiovisual dual attention task (Barutchu et al., 2019). Thus, the children who perceived the McGurk effect more frequently were better able to simultaneously attend to information in the auditory and visual modalities. Perception of the McGurk effect, or the integration of auditory and visual speech is related to measures of academic readiness and dual attention in school-aged children.

The McGurk Task: Summary & Future Directions

Table 3 displays a summary of all studies assessing intersensory processing using the McGurk task. Research assessing perception of the McGurk effect has demonstrated that infants, pre-school, and school-aged children perceive the McGurk effect and thus, show audiovisual integration. However, children perceive the McGurk effect less often than adolescents and adults, likely due to auditory dominance in early-childhood that shifts across development to adult-like levels of visual dominance by middle-childhood. Given that the McGurk effect depends on visual lip movements, this shift from auditory to visual dominance may in part explain why the susceptibility to the McGurk effect increases across development. The McGurk effect appears to be stronger for English-speaking compared to Japanese-speaking children, possibly due to cultural differences in the promotion of visual and auditory information in the language environment. Infants at high risk for autism and children with autism shower poorer intersensory processing of the McGurk effect than TD children. In contrast, for children with language delays and disabilities, perception of the McGurk appears to vary as a function of the specific developmental disability. Finally, greater perception of the McGurk effect is related to

greater academic achievement and divided audiovisual dual attention in school-aged children.

There are a number of future research directions for assessing intersensory processing using the McGurk task. First, research should examine whether young infants perceive the McGurk effect. Research using the McGurk task has found evidence of detecting the McGurk effect in 4- and 5-month-old infants (Burnham & Dodd, 1996, 2004; Desjardins & Werker, 2004; Rosenblum et al., 1997), but habituation studies have demonstrated that much younger infants are capable of discriminating audiovisual information (Bahrick, 1992; Flom et al., 2018). Second, the developmental studies conducted testing perception of the McGurk effect have all been cross-sectional thus far (Dupont et al., 2005; Hirst et al., 2018; Massaro, 1984; McGurk & MacDonald, 1976; Tremblay et al., 2007). Future research should incorporate longitudinal designs to assess how individual children change across development in perceiving the McGurk effect. Third, future research could assess the perception of the McGurk effect in a variety of languages (other than Japanese) to explore the role that the language environment plays in intersensory processing of incongruent audiovisual speech. Fourth, more research could be conducted in infants at risk for autism and children with autism to identify atypical developmental progressions and children at risk for intersensory processing delays. This direction for future research also calls for longitudinal designs probing a wider range of ages in both typically and atypically developing children. Further, research could examine whether the perception of the McGurk effect in children with language delays/disabilities is dependent on the specific delay/disability and the degree of language proficiency they exhibit. Finally, future research should examine whether

greater perception of the McGurk effect is related to greater social, cognitive, or language skills.

Speech-In-Noise Task

Method & Research Questions

The speech-in-noise task (O'Neill, 1954; Sumby & Pollack, 1954) addresses a participant's ability to use visual information to aid in perceiving auditory speech when it is degraded by noise. This method features a video display of a person speaking with added noise (e.g., white noise, multi-talker babble) in the background presented at multiple signal-to-noise ratios (SNRs) where the power of the speech signal is compared with the power of the noise. This ratio is often expressed in decibels, with a ratio greater than 0-dB indicating more signal than noise. The general question addressed by the speech-in-noise task is the extent to which the audiovisual redundancy provided by lip movements during speech facilitates perception of auditory speech in noise (as compared with auditory-only or visual-only speech).

The speech-in-noise task shares some similarities with the paradigms reviewed previously. Similar to the McGurk task, the speech-in-noise task assesses the influence of visual information on auditory speech perception. Similar to the McGurk task (and in contrast with the intermodal preference and habituation methods), it is also designed to address a narrowly focused research question concerning the influence of visual speech on the ability of children to perceive and understand auditory speech. Also unlike the intermodal preference and habituation methods, it typically requires a verbal response from the participant and thus the majority of research is conducted with verbal children. It is thought that when both the audible and visible speech streams are present, speech

processing is most efficient, and when the auditory stream is degraded, the visual speech information aids in accurate processing. In the following section, we review research findings of intersensory processing as assessed by the speech-in-noise task. We also include a comprehensive review summarizing all studies assessing intersensory processing of faces and voices using the speech-in-noise task in young children (see Table 4).

Research Findings

In contrast with the large body of research focusing on adults (e.g., Erber, 1969; Grant & Seitz, 2000; Macleod & Summerfield, 1987; Nath & Beauchamp, 2011; Sumby & Pollack, 1954), very little research has been conducted with children using the speech-in-noise task (see Table 4). Children ranging in age from 5 to 14 years showed significant audiovisual gain (e.g., difference in performance between audiovisual and auditory-only conditions) when identifying mono-syllabic words presented at six different levels of pink noise (i.e., no noise, -3, -6, -9, -12, -15 dBs; Ross et al., 2011). However, children do not show as much gain as do adults. At 5 to 7 years, they showed 27.8% audiovisual gain compared to 52.7% in adults. Audiovisual gain increased linearly across age and child performance reached adult-like levels at approximately 12 to 14 years of age. Children make increasing use of visual information to help identify speech in noise across age and reach adult levels by 12 to 14 years of age.

Three studies have focused on speech-in-noise perception in children with developmental disabilities. In a sample of children with autism, 6- to 18-year-olds showed lower accuracy in identifying whole-words and phonemes presented at four SNRs (0, -6, -12, -18 dBs) compared to TD children, who showed increasing gains in

audiovisual accuracy as the SNR declined (Stevenson et al., 2017). In a sample of hearing-impaired children, 9- to 12-year-olds required a greater SNR (i.e., a positive ratio, indicating more signal than noise) to identify words presented in noise than TD children (Erber, 1971). Finally, children with language learning impairment (LLI) and TD children showed increasing accuracy across 4 to 11 years of age in the identification of target words from sentences presented in noise, with no significant difference between them (Knowland et al., 2016). Children with autism and hearing-impairments, but not those with LLI, show reduced benefit from visual speech information compared to TD children and adults.

Speech-in-Noise Task: Summary & Future Directions

Table 4 presents a summary of the small body of research using the speech-in-noise task to assess intersensory perception in children. Children benefit from visual information for identifying words presented in noise, but show lower levels of audiovisual gain than do adults. Children with autism and hearing-impaired children (but not those with LLI) are less accurate at perceiving speech-in-noise than TD children.

There are a number of possible future directions for research using the speech-in-noise task. First, studies have focused on children who have good verbal skills and there are no studies with children under the age of five. Thus, the extent to which infants benefit from visual information during early speech perception is not known. However, we do know from the other methods reviewed earlier, that infants are already skilled at perceiving amodal information including temporal synchrony, prosody, affect, rhythm, and tempo, uniting auditory and visual speech (Bahrick et al., 2019; Flom & Bahrick, 2007; Lewkowicz, 2003, 2010; Walker, 1982). Detection of these amodal properties

likely underlies the gains seen from visual speech for perceiving speech in noise, and thus it would be expected that infants would show gains just as do children. Further, research is also needed to identify the specific intersensory skills that underlie gains in speech perceptibility from visual speech. Future research should thus examine whether infants show gains from visual speech in identifying syllables in the context of noise using methods such as habituation or intermodal preference. Second, thus far, there is only cross-sectional evidence that audiovisual gain for speech-in-noise perception improves with age (Knowland et al., 2016; Stevenson et al., 2017). Future research should use a longitudinal approach to establish developmental trajectories for speech-in-noise perception in TD infants and children so that atypical developmental patterns can be detected in early development. Finally, in conjunction with this, future research could examine whether greater audiovisual gains result in better developmental outcomes.

The Eye-Tracking Method

Method & Research Questions

The eye-tracking method has emerged recently with the advent of new technology as a method for assessing selective attention to different areas of a visual display. It records across time, the specific areas of an image or event the participant focuses on (e.g., the mouth vs. eye region of a speaking face). Gaze direction is determined by the reflection of infrared light sources on the eye(s) projected from the eye-tracker. A calibration process occurs prior to data collection in order to provide an external frame of reference for the participant's gaze behavior. It involves recording information about the participant's pupil(s) and corneal reflection(s) for fixations at known locations on the screen (Oakes, 2010, 2012). The eye-tracking system then uses information obtained

from the calibration process to determine the point-of-gaze across exploratory time for a videotaped presentation of image or event. (3-D eye-trackers are also used but typically not for assessing detection of audiovisual relations. They are more specialized for real-time visual attention or visual-motor behaviors). Thus, eye-tracking allows one to obtain automated measures of gaze location across exploratory time, typically categorized into areas of interest (AOIs; mouth, eye region) with no need for human data coding (Hessels & Hooge, 2019). This review focuses on infant attention to speaking faces using eye-tracking measures used to examine which parts of the face infants fixate during audiovisual speech (e.g., AOIs; eyes, mouth, central region). It is assumed that the mouth area depicts greater synchrony information than other areas of the face during audiovisual speech (Lewkowicz & Hansen-Tift, 2012) and thus looking to this area should correlate with face-voice synchrony detection. Therefore, the general purpose of the eye-tracking methods reviewed here is to characterize the differences in selective attention to parts of the face, typically in synchronous audiovisual speech, as a function of the vocal information presented (infant vs. adult directed; native vs. non-native speech) and under what conditions infants attend to specific areas of the face that depict the greatest audiovisual synchrony.

In contrast with the use of human observers or coders for collecting data from the intermodal preference and habituation methods, eye-tracking can provide much greater detail about gaze direction relative to areas of a video display and how it is distributed across time. Human observers are quite reliable at determining gaze direction to two screens or visual recovery to one screen and thus eye-tracking is most useful when research questions involve assessing the distribution of selective attention to specific

AOIs. Eye-tracking has the disadvantage relative to the other methods, however, that it can involve significant data loss due to infant movement, attention off-screen, closing the eyes, or because the eye-tracker is unable to detect the eyes, cornea, or pupil reflection (Hessels et al., 2015). In the following section, we review selective research findings of intersensory processing as assessed by the eye-tracking method. We also include a comprehensive review summarizing all studies assessing intersensory processing of faces and voices using eye-tracking in infants and young children (see Table 5).

Research Findings

Selective Attention to Specific Areas of Speaking Faces and Change Across Development. One primary focus of research using eye-tracking has been to examine selective attention to specific areas (typically the eyes vs. mouth) of a speaking face and how it changes across development (see Table 4). In one of the earliest studies, Lewkowicz and Hansen-Tift (2012) assessed infant visual attention to monologues spoken in a native language (i.e., English) and a non-native language (i.e., Spanish) in either infant-directed or adult-directed speech across the first year of life. Regardless of whether speech was infant-directed or adult-directed, infants showed greater overall looking to the mouth of the speaker. When viewing the monologue in their native language, 4-month-olds looked longer to the eyes, 6-month-olds looked equally to the eyes and the mouth, 8- and 10-month-olds looked longer to the mouth, and 12-month-olds looked equally at the eyes and mouth. The same pattern of looking was found when infants viewed the monologue in a non-native language, with the exception that 12-month-olds looked more to the mouth. This extended looking to the mouth was assumed to reflect perceptual narrowing and growing expertise for the native language. That is, it

was hypothesized that infants remained focused on the mouth (where the greatest amount of intersensory redundancy is assumed to be available) to aid audiovisual speech processing for the non-native language. The attentional shift to the mouth at 8 months has also been replicated in monolingual Spanish infants (Pons et al., 2015, 2019) and when the auditory and visual speech streams of the monologues are asynchronous (Hillairet de Boisferon et al., 2017). Thus, a conclusion that has emerged from this research is that infants focus preferentially on the mouth area when first learning the sounds of the language. Given that data are analyzed at the group level (and are not correlated with individual differences in language skills) data provides indirect evidence for this view. Together, these studies suggest that selective attention to speaking faces follows a U-shaped developmental pattern, with greater attention to the eyes around 4 months, then to the mouth, and then trending back to the eyes around the first birthday. Further, at 12 months of age, infants differ in their attention to the speaker's mouth as a function of language familiarity.

A number of studies examining selective attention to speaking faces have found that attention to the mouth relative to the eyes increases across development. Attention to the mouth of a speaking face increased from 6 to 12 months of age (Tsang et al., 2018). It was also found to increase from 5 months to 5 years of age, but with no significant differences between preference for the eyes versus the mouth at 12 months of age (Morin-Lessard et al., 2019). When shown a woman speaking at a table with objects in front of her at 6, 9, and 12 months of age, infants looked more to the woman's face (compared to the objects) at all ages, and more to the mouth than the eyes at 9 and 12 months of age (Tenenbaum et al., 2013). Infants 12 and 18 months of age looked

significantly longer to the mouth of a speaking face than infants 6 months of age (Imafuku et al., 2019). Finally, when presented with native and non-native speech monologues in infant-directed and adult-directed speech, infants 14 months of age looked longer to the mouth during infant-directed speech, but not adult-directed speech, regardless of the language spoken. However, infants 18 months of age looked longer to the mouth regardless of language spoken and whether the speech was infant- or adult-directed (Hillairet de Boisferon et al., 2018). These studies converge with findings from Lewkowicz & Hansen-Tift (2012) in that infants pay a great deal of attention to the mouth of a speaking face. However, these studies did not find the same developmental pattern of shifting attention to the eyes relative to the mouth for native speech at the end of the first year. According to these studies, selective attention to the mouth of a speaking face relative to the eyes increases across the first year and a half of life, regardless of whether the speech is in a native or non-native language.

In contrast, a few studies have found attention to the eyes of a speaking face relative to the mouth area increases across the first year. For example, a study using the McGurk task found that attention to the mouth did not increase across 6 to 9 months of age for the congruent speech, but increased for the incongruent, nonredundant speech (e.g., visual “ba” with auditory “ga” that produces a combined response that cannot be fused into a single percept; Tomalski et al., 2013). When presented with a video of a woman waving and saying “hey baby”, infants 3 to 4 months of age looked equally to the eyes and mouth, while infants 9 months of age looked more to the eyes (Wilcox et al., 2013). Together these studies found that attention to the eyes relative to the mouth of a speaking face increases across the first year. Differences in findings may be due to the

stimuli having less linguistic content (compared to other studies using monologues). Thus, attention to specific areas of a speaking face appear to depend on a variety of factors and data are somewhat inconsistent across studies using different stimulus events and contexts.

Selective Attention to Speaking Faces in Monolingual Versus Bilingual Children. Another focus of research using eye-tracking has been to examine whether selective attention to specific areas (i.e., eyes, mouth) of a speaking face differ between monolingual and bilingual language learning infants and children (see Table 5). When presented with women speaking in a native and a non-native language, monolingual 4-month-old infants looked longer at the eyes than the mouth, but bilingual 4-month-old infants looked equally at the eyes and mouth. At 8 months, both monolingual and bilingual infants looked longer to the mouth than the eyes for both native and non-native speech. At 12 months, monolingual infants looked equally to the eyes and mouth for native speech and longer at the mouth than the eyes for non-native speech, whereas bilingual infants looked longer to the mouth than the eyes for both native and non-native speech. Thus, bilingual infants showed an earlier attentional shift to the mouth and appeared to take greater advantage of the intersensory redundancy provided by the mouth than monolingual infants did (Pons et al., 2015). In contrast, attention to the mouth of a speaking face did not differ between monolingual and bilingual infants at 6 (Tsang et al., 2018), 12 (Morin-Lessard et al., 2019; Tsang et al., 2018), and 15 months (Fort et al., 2017) when presented with a native language (Fort et al., 2017; Tsang et al., 2018) or both a native and non-native language (Morin-Lessard et al., 2019). The pattern of selective attention to the eyes and mouth of women speaking for monolingual versus

bilingual infants remains unclear. These differences may be due to differences in stimuli, languages presented, definitions of bilingualism, etc., and remain a topic of future research.

Selective Attention to Speaking Faces in Infants and Children at Risk for or Displaying Developmental Disabilities. Eye-tracking has also been used to examine selective attention to speaking faces in infants and children at risk for or displaying developmental disabilities. This body of research has assessed whether these children pay attention to similar areas of a speaking face compared to TD infants and children (see Table 5). One study focusing on infants at risk for autism demonstrated that at 6 months of age, infants who were later diagnosed with autism looked to the inner features of a speaking face less than TD infants (Shic et al., 2014). Two studies have focused on infants born pre-term. Using the intermodal preference method, infants born pre-term looked at the mouth of a speaking face (congruent or incongruent) significantly longer at 18 months than they did at 6 or 12 months, whereas infants born full-term looked significantly longer at the mouth at 12 and 18 months compared to 6 months. Further, at 18 months of age, infants born pre-term spent less time looking to the mouth of a speaking face (congruent or incongruent) than TD infants (Imafuku et al., 2019). When presented with a woman speaking in native and non-native languages, 8-month-old infants born pre-term did not show differential looking to the eyes versus the mouth for the native or non-native languages, whereas 6- and 8-month old full-term infants looked more to the eyes when hearing the native language and more to the mouth when hearing the non-native language (Berdasco-Muñoz et al., 2019). Infants born pre-term and those later diagnosed with autism display different patterns of selective attention to faces

speaking in across a variety of conditions (native vs. non-native language, congruent vs. incongruent speech side-by-side) across 6 to 18 months of age.

Finally, one study has focused on children with SLI. It demonstrated that children 5 to 9 years of age with SLI looked equally to the eyes and mouth of a speaking face, whereas typically developing children looked more to the mouth. For only the children with SLI, those who had lexical-syntactic deficits (related to forming sentences with words) spent more time looking to the mouth compared to the children with SLI who had phonological-syntactic deficits (related to forming sounds, such as phonemes and syllables; Pons et al., 2018). Children with SLI who had phonological-syntactic deficits appeared to drive the pattern observed in the overall SLI group compared to the TD children. Children with SLI show a different pattern of selective attention to faces depicting continuous speech than their TD peers in early to middle childhood. However, this difference depends on the nature of the SLI. Regardless, reduced looking to the mouth of a speaking face in children with SLI may underlie their deficits in integrating and processing audiovisual speech compared to TD children.

Selective Attention to Speaking Faces and Relations with Developmental Outcomes. A number of studies have examined how selective attention to specific areas (i.e., eyes, mouth) of a speaking face are related to developmental outcomes (see Table 5). Some of these studies have examined concurrent relations between selective attention to speaking faces and language skills. For example, looking to the mouth of woman speaking was associated vocabulary size concurrently at 9, 12, 14, 18, and 24 months of age (Morin-Lessard et al., 2019). In contrast, there was no significant relation between looking to the mouth of a woman speaking and expressive vocabulary at 18 months of

age (Hillairet de Boisferon et al., 2018). At most ages, selective attention to the mouth of a woman speaking continuously is related to concurrent child vocabulary size. These studies provide the first direct evidence of a link between mouth looking and language skills.

A number of studies have also assessed prospective relations between selective attention to speaking faces and language outcomes. For example, at 6 months of age, infants who spent a greater amount of time fixating their mother's mouth (as compared with her eyes) in an interaction had greater expressive vocabulary scores at 24 months of age. Further, their vocabulary was more advanced by 4 months than the vocabulary of the infants who looked less at the mouth relative to the eyes (Young et al., 2009). When presented with a woman speaking about objects in front of her, a greater proportion of attention to the mouth relative to eyes at 12 months predicted unique variance in expressive vocabulary at 18 and 24 months (Tenenbaum et al., 2015). Selective attention to the mouth of a woman speaking is prospectively related to language outcomes in a variety of contexts (familiar vs. unfamiliar women, interaction with vs. observation of woman speaking), indicating evidence of links between this measure and later language skills.

Some studies have examined relations between selective attention to speaking faces and language skills in monolingual versus bilingual language learning infants. A greater proportion of attention to the mouth relative to eyes of a speaking woman at 6 months of age was associated with greater expressive vocabulary at 12 months of age in both monolingual and bilingual children (Tsang et al., 2018). However, in a different study, greater looking to the mouth of a woman speaking was related to greater

concurrent expressive vocabulary in monolingual children, but not bilingual children at 9, 12, 14, 18, and 24 months of age (Morin-Lessard et al., 2019). Overall, studies treating mouth versus eye looking as an individual difference measure for predicting language outcomes provides direct evidence that this variable is a meaningful index of language learning (likely including attention to audiovisual speech and intersensory processing). Selective attention to the mouth area was related to concurrent and later expressive vocabulary in monolingual children. In contrast, evidence is inconsistent across studies with respect to bilingual children.

Some studies have assessed relations between selective attention to women speaking in congruent versus incongruent audiovisual speech and language skills. When shown women telling stories in congruent and incongruent (video played backwards) speech, greater time spent looking at the mouth of a woman speaking at 6 months was related to greater receptive vocabulary at 12 months, regardless of whether the speech was congruent or incongruent (Imafuku & Myowa, 2016). In contrast, when presented with incongruent audiovisual speech, greater looking to the eyes of a woman speaking syllables during the incongruent speech at 6 to 9 months of age was associated with greater auditory comprehension of language at 14 to 16 months, whereas greater looking to the mouth during incongruent speech was associated with poorer auditory comprehension of language (Kushnerenko et al., 2013). This finding appears to be inconsistent with the hypothesis that mouth looking reflects detection of audiovisual face-voice synchrony. Selective attention to the mouth over the eyes during speech thus far appears to be the most promising measure of detection of audiovisual face-voice relations, but findings are inconsistent. In studies successfully predicting language

outcomes, selective attention to the mouth was related to concurrent and later expressive vocabulary. However, when the audiovisual speech was asynchronous, greater auditory comprehension of speech was found for both mouth and eye looking, and thus, it remains unclear if mouth looking reflects the detection of audiovisual face-voice synchrony.

Finally, one study has assessed concurrent relations between selective attention to speaking faces and social skills. It is hypothesized that looking to the eyes promotes infant responsiveness to social, deictic, and referential information provided by the eyes, whereas looking to the mouth aids in disambiguating speech sounds. Consistent with this view, at 12 months of age, greater proportion of attention to the eyes of a woman speaking in a native or non-native language was associated with greater social skills involving social interactions and joint attention processes (Pons et al., 2019). Greater selective attention to the eyes of a woman speaking continuously is related to greater concurrent social skills.

The Eye-Tracking Method: Summary & Future Directions

Table 5 displays a summary of all studies assessing intersensory processing with eye-tracking. Research using eye-tracking has demonstrated that at 12 months of age, selective attention to the eyes versus mouth of a woman speaking depends on whether the speech is in a native or non-native language, but prior to this, infants show consistent patterns of attention to the eyes versus mouth regardless of language. Some studies have found that attention to the mouth increases across the first year and a half of life, whereas other studies presenting less linguistic content have found that attention to the eyes increases across the first year. Thus far, the basis for differences in attention to the eyes versus mouth of a woman speaking for monolingual versus bilingual infants remains

unclear. Infants born pre-term and children with SLI appear to have reduced attention to the mouth of a speaking person compared to their TD peers. Greater attention to the mouth (relative to the eyes) of a speaking face is related to greater language outcomes. Thus, it is evident that reduced attention to the mouth of a speaking face can have a negative cascading effect on later language.

One important question for this area concerns the extent to which findings from eye-tracking studies actually reflect detection of intersensory audiovisual relations (i.e., temporal synchrony or nested amodal relations). Unlike the methods reviewed previously that each provide direct evidence of intersensory processing skills in groups of infants via careful manipulation of stimulus contrasts, the basis for preferential looking to the mouth over eyes during speech is thus far not clear. It is likely that the mouth area provides a greater amount of temporal synchrony between the sounds of speech and movements of the face than other areas of the face, though this would vary across stimulus event and speech types. There has been little attempt to assess the degree of synchrony in different displays, but studies that have presented stimulus conditions that presumably differ in amount of synchrony shown in the mouth area (infant directed vs. adult directed speech) and in a few cases, synchronous versus asynchronous speech, have not consistently found greater mouth looking in conditions that provide greater temporal synchrony (Hillairet de Boisferon et al., 2018; Imafuku & Myowa, 2016; Tomalski et al., 2013).

The use of selective attention to the mouth as an individual difference measure capable of predicting developmental outcomes, particularly language skills, is a more promising approach. It has provided direct evidence linking selective attention to the mouth with language skills. Thus far, these studies provide fairly consistent evidence

that selective attention to the mouth area predicts both concurrent and later expressive vocabulary across the first two years of life in monolingual language learning children (Morin-Lessard et al., 2019; Tenenbaum et al., 2015; Tsang et al., 2018; Young et al., 2009; but see Hillairet de Boisferon et al., 2018 for an exception) but evidence for bilingual language learning children is inconsistent. This body of research indicates that selective attention to the mouth is a meaningful variable for predicting later language outcomes across a variety of conditions. However, the specific information detected by preferential fixation of the mouth area has not yet been a topic of systematic investigation.

There are several future directions for studies examining intersensory processing of speaking faces with eye-tracking. First, the developmental trajectory of selective attention to the eyes versus mouth of a speaking face remains unclear. Using a longitudinal approach, research should establish the typical trajectory of looking to the eyes versus the mouth of a speaking face in both monolingual and bilingual populations. This will not only clarify what a typical trajectory (versus an atypical trajectory) looks like, but will also clarify differences between monolinguals and bilinguals. Further, studies should be conducted to directly link eye-tracking measures such as selective attention to the mouth to detection of temporal synchrony (or other amodal audiovisual relations). Such studies could systematically compare conditions of synchronous audiovisual speech with asynchronous audiovisual speech and unimodal visual speech to determine if greater selective attention to the mouth is evident in the conditions of greater synchrony. Further, additional eye-tracking variables that relate to detection of audiovisual relations could be identified. Fourth, unlike the other paradigms, an

individual difference approach and longitudinal studies have already been conducted using the eye-tracking method. Future research should expand the use of these approaches in order to establish the range of conditions under which selective attention to the mouth versus eyes of a speaking face is related to better language versus social outcomes.

New Individual Difference Approaches

The paradigms reviewed above that were available during the early development of the field of intersensory processing (1970s through 2000) including intermodal preference and habituation methods, the McGurk task, and the speech-in-noise task (all but the eye-tracking method which emerged more recently with the advent of eye-tracking technology), have been designed for and conducted using a group differences approach to assess intersensory processing. The research generated from these paradigms has produced a wealth of information about infant intersensory processing skills, providing details about the specific intersensory processing skills of groups of infants at specific ages and under specific conditions. This body of research has contributed to the development of theory and has generated a valuable knowledge base about the nature of intersensory processing skills across infancy. However, these methods were not designed for use as individual difference measures of intersensory processing, typically having poor reliability (Colombo et al., 2004), presenting few trials and thus having a coarse grain of analysis, and are used primarily in cross-sectional research designs. Other areas of research including studies of language and cognitive functioning have benefitted from the use of a range of individual difference measures assessing skills of individual infants relative to one another. This allows researchers to assess developmental trajectories, and

predictive relations between early developing skills and later outcomes. The lack of reliable individual difference measures in the area of intersensory processing has thus limited the nature of research questions that could be addressed. To address this need, Bahrick and colleagues have developed the first two individual difference measures designed to assess specific intersensory processing skills across infancy and childhood. (Bahrick et al 2018a, 2018b).

The Multisensory Attention Assessment Protocol (MAAP; Bahrick, Todd et al., 2018) assesses three "multisensory attention skills" (i.e., intersensory processing, sustained attention, and speed of shifting/disengaging) to audiovisual events (in contrast with prior paradigms which have typically assessed a single skill), allowing researchers to examine relations among specific attention skills deployed in the context of dynamic audiovisual events and relations with later outcomes. The Intersensory Processing Efficiency Protocol (IPEP; Bahrick, Soska et al., 2018) is a fine-grained and more comprehensive measure of just intersensory processing. It assesses both speed and accuracy of intersensory processing skills (in contrast, prior methods have typically assessed only accuracy of intersensory processing). Both measures assess intersensory processing skills at a sufficiently fine-grained level to address novel research questions regarding the performance of individual children relative to one another across age. These include using both longitudinal and cross-sectional designs to characterize developmental trajectories of intersensory processing skills, predict developmental outcomes from intersensory processing skills, and explore models of developmental pathways between basic intersensory processing skills and later developmental outcomes that rely on this foundation.

Overview of the MAAP and IPEP

The MAAP is a three-screen video-based procedure that assesses attention to dynamic, audiovisual social and nonsocial events. It presents two lateral events (right and left sides of the display) depicting two different social events (e.g., two different women telling a different story; see Figure 1 or two nonsocial events, not pictured here) across a number of short trials (24 social and 24 nonsocial). The movements of one of the lateral events are synchronous with its natural soundtrack, while the movements of the other are asynchronous. A central stimulus event (i.e., morphing geometric shapes) is presented in the middle screen for half of the trials to provide an additional source of competing stimulation (high competition trials; see bottom, Figure 1), and there is no central stimulus event during the other half of the trials (low competition trials; see top, Figure 1). The measure and general question assessed by the MAAP, similar to that of the intermodal preference method, is the proportion of total looking time (PTLT) infants look to the sound-synchronous event (as a function of looking to both the synchronous and asynchronous events; i.e., synchrony detection). This can be used as an individual difference variable or similar to the intermodal preference method, one can assess whether the group of infants shows significant evidence of detecting amodal audiovisual relations (including global synchrony, rhythm, tempo, intensity changes common to the face and voice). The MAAP also assesses overall interest (attention maintenance) in the dynamic social events in the context of a soundtrack appropriate to one event, as well as their speed of shifting/disengaging to look to the dynamic social events, both considered multisensory attention skills (but not indicative of intersensory processing).

In contrast, the IPEP is a fine-grained measure of just intersensory processing skills (speed and accuracy of matching) in the context of dynamic, audiovisual social and nonsocial events. It features six concurrent events arranged in two rows of three accompanied by a single soundtrack (see Figure 2, for social events) presented across a large number of short trials (24 social and 24 nonsocial). The social events depict six different women telling each telling a different story. For each trial, the movements of one the events (target event) are synchronous with the soundtrack, while the movements of the other five events (distractor events) are asynchronous. The measures and general question addressed by the IPEP include how quickly infants can locate the sound-synchronous event (speed; latency to fixate), how long (duration) they fixate the sound-synchronous target event as a function of their looking to the five asynchronous distractor events (PTLT), and how frequently they fixate the sound-synchronous event. Each of these measures can be used as both individual differences variables or assess performance at the group level. Similar to the MAAP, the PTLT to the synchronous event can assess whether the group of infants shows significant evidence of detecting amodal audiovisual relations (including global synchrony, rhythm, tempo, intensity changes common to the face and voice) as well as how quickly.

Although both the MAAP and IPEP assess detection of audiovisual synchrony, similar the intermodal preference method, there are important differences. Both the MAAP and IPEP include competing stimulation (a central distractor event in the MAAP and five distractor events in the IPEP) simulating the noisiness of the natural environment. Further, the MAAP and IPEP are based on a relatively large number of short trials, that can be averaged to provide a relatively stable score for an infant, serving

as an individual variable. These scores have good to very good reliability (Bahrick, Todd, et al., 2018). Thus, when used as an individual difference variable, the MAAP and the IPEP can characterize how infants perform relative to other infants and can assess relations between basic intersensory processing skills and those in other domains, such as language and cognitive functioning. Therefore, the MAAP and IPEP can address questions regarding skills of groups of infants, similar to the intermodal preference method, but can also be used to address a variety of important questions that cannot be addressed with any of the group differences approaches.

A rich range of important questions can now be addressed by the fine-grained assessment of individual differences in intersensory processing afforded by the MAAP and IPEP. For example, models of developmental growth can be constructed to identify typical and atypical trajectories of intersensory processing across the first few years of life. Individual differences in intersensory processing can be correlated with individual differences in developmental outcomes to assess our ability to predict functioning in other domains such as language and cognition from early developing, basic, intersensory processing skills. Multiple regressions can assess the unique variance intersensory processing skills contribute to these later outcomes, controlling for other well-known predictors in those domains. Pathways of development can also be tested with structural equation models to discern cascades from early intersensory processing skills to later developmental capabilities that rely on these skills. In the following section, we review recent research findings that address some of these new questions in the domain of intersensory processing of social events as assessed by the MAAP and the IPEP.

Research Findings from the MAAP and IPEP

Research has recently begun to examine intersensory processing of social events using the MAAP. For example, intersensory processing of social events (proportion of looking to the sound-synchronous social event) has been found to show significant linear growth across 3 to 12 months of age (Todd et al., 2017). Individual differences in intersensory processing have also been related to individual differences in child language outcomes. Greater intersensory processing of social events was correlated with greater receptive and expressive vocabulary in 2- to 5-year-old children. Further, results of a structural equation model revealed that sustained attention to social events predicted intersensory processing of social events, which in turn predicted receptive and expressive vocabulary in children of this age (Bahrick, Todd et al., 2018). In another study, Edgar, et al. (under review) assessed the contribution of intersensory processing to language outcomes in the context of other well-known predictors of language including parent language input and SES. They found that intersensory processing of social (but not nonsocial) events in the presence of the distractor at 12 months of age predicted child quality and quantity of speech production, as well as expressive vocabulary at 18 and 24 months of age, while controlling for parent language input (quality and quantity) and SES. Moreover, in many cases intersensory processing accounted for more unique variance in child language outcomes than parent language input and SES combined (Edgar et al., under review).

Research has also recently examined intersensory processing of social events using the IPEP to reveal fine-grained differences in both speed and accuracy of intersensory processing. For example, accuracy of intersensory processing of social

events (proportion of time fixating the sound-synchronous woman) showed significant growth across 3 to 36 months of age (Todd & Bahrick, 2020). Individual differences in speed and accuracy of intersensory processing have also been related to individual differences in child language outcomes. At 6 months of age, accuracy of intersensory processing (but not speed) for social (but not nonsocial) events predicted child language outcomes at 18, 24, and 36 months, over and above the contribution of well-known predictors of parent language input and SES (Edgar et al., under review). Thus, research using the MAAP and IPEP is beginning to determine the trajectory of intersensory processing of social events and establish direct links between early intersensory processing skills and later language outcomes.

The MAAP and IPEP: Summary & Future Directions

The MAAP and the IPEP offer the first and thus far, only individual difference measures designed to assess intersensory processing of audiovisual events in preverbal infants and children. These measures can assess intersensory processing at a sufficiently fine-grained level for predicting outcomes in other domains. The availability of these measures opens the door to assessing a host of important questions that the group difference approaches are not designed to address, including relations between intersensory processing skills and later outcomes, developmental pathways between intersensory processing skills and later outcomes, and well as developmental growth and change in these skills. Research using the MAAP and IPEP is just beginning to address these questions. Findings demonstrate developmental improvement of intersensory processing of social events, as well as direct links between individual differences in early intersensory processing skills and individual differences in later child language outcomes.

Future research in our lab will establish the developmental trajectories of various intersensory processing of social events across the first 6 years of life. An important challenge for future research will be to assess how intersensory processing cascades to later developmental outcomes such as language, cognitive and social functioning. This research has the potential to transform theory, our understanding of mechanisms and processes leading to language outcomes, and provide applications for identifying the atypical development of intersensory processing and infants at risk for later impairments in language and cognitive functioning.

General Conclusions and Future Directions

Five decades of research demonstrate that infants and children have a wide and rich range of intersensory processing skills in the domain of social events. This finding is informed by a variety of measures and paradigms designed to assess intersensory processing skills, each with specific research questions and somewhat different conclusions that can be drawn. The largest body of research has been generated by the intermodal preference method. Research using this method has demonstrated that infants detect and match faces and voices of people speaking based on a variety of amodal (including temporal synchrony, spectral information, prosody, affect, gender, and age) and modality-specific relations (including the specific appearance of the face and sound of the voice uniting the sights and sounds of speech). The habituation method has demonstrated that infants discriminate a change in the face and voice of a speaking person on the basis of amodal audiovisual relations including temporal synchrony, rhythm, affect, prosody, and arbitrary relations (e.g., face-voice pairings, syllable-object

relations). Research using the McGurk task has demonstrated that infants and children perceive the McGurk illusion (requiring fusing incongruent audible and visible speech information), providing evidence of audiovisual integration for a specific type of stimulus. The speech-in-noise task has demonstrated that children benefit from the intersensory redundancy provided by lip movements during audiovisual speech, and that performance improves to near adult levels by 12 to 14 years of age. These methods, designed for group-level analyses have provided a significant and substantial body of data documenting specific intersensory processing skills in infants and children of various ages and under a variety of conditions for a range of stimulus events.

Research using the eye-tracking paradigm for assessing attention to audiovisual speech emerged more recently (since approximately 2009) as a result of the emergence of eye-tracking methodology, and has been used for both group-level and individual difference approaches. Group-level studies show that infants and children, deploy selective attention to specific areas of the face (i.e., mouth as compared with eyes) that provide intersensory redundancy, however the basis for selective looking to the mouth has not yet been clearly linked to intersensory processing skills. Importantly, research has also demonstrated that individual differences in selective attention to the mouth predict concurrent and later expressive vocabulary. Finally, the individual differences approach assessing specific intersensory processing skills across multiple short trials (using the MAAP and IPEP) has opened the door to asking a host of new questions about developmental processes. These include the ability to map developmental trajectories of intersensory processing skills for typical and atypical development, predict concurrent and future outcomes, including language skills, and derive models depicting pathways

between early developing intersensory processing skills and later outcomes. Although the different methods and stimulus events used present a challenge for making comparisons and drawing conclusions across studies, they also highlight a number of convergent findings across measures and paradigms, suggesting general principles of development, patterns of atypical development, as well as important gaps in the body of research for future studies to address.

Convergent Findings Across Paradigms and Future Directions

The Foundational Role of Temporal Synchrony Detection. Research across paradigms supports the conclusion that temporal synchrony serves as an important foundation for matching faces and voices in early development. Studies using the intermodal preference and habituation methods demonstrate that infants can detect changes in and match faces and voices of people speaking on the basis of temporal synchrony, even as neonates. Moreover, the McGurk task demonstrates the powerful role that temporal synchrony plays in integrating audible and visible speech. Findings demonstrate that an incongruent audible and visible speech sound spoken in temporal synchrony can be fused and perceived as an emergent, unified percept as early as 4 months of age. Thus, there is clear evidence across several paradigms that temporal synchrony is a fundamental basis for matching faces and voices.

Detecting Higher-Level Amodal Properties. Research also supports the conclusion that detecting temporal synchrony provides a basis for detecting nested levels of audiovisual relations specifying properties such as prosody, affect, age and gender of speaker, and spectral information in speech sounds. Findings from the intermodal preference and habituation methods indicate that audiovisual matching of prosody,

spectral information, arbitrary syllable-object relations, and at certain ages, affect and rhythm, occurs in the context of face-voice synchrony and not in its absence. Findings demonstrating detection of the McGurk illusion illustrate the powerful role of temporal synchrony in promoting fusion of incongruent audible and visible speech information, creating perception of a novel speech sound. Results of the speech-in-noise task indicate that adding temporally synchronous visual information to auditory speech can lead to improved perception of speech sounds. Findings across several paradigms converge to support the role of temporal synchrony in detecting nested levels of audiovisual relations.

Relatedly, the principle of intersensory facilitation predicted by the IRH has also been tested and supported by research across several paradigms. The intermodal preference and habituation methods converge to provide evidence that amodal information for properties such as affect, prosody, rhythm, and tempo are detected best in audiovisual synchronous stimulation as compared with asynchronous or unimodal (auditory of visual) stimulation. This has been clearly demonstrated in the domain of nonsocial events and also extended to perception of faces and voices. Given this knowledge base, research can address more specific questions about intersensory facilitation. For example, can synchrony detection be improved through training and generalized to other social contexts and events? How can the principle of intersensory facilitation be applied in multimodal educational contexts to improve language, social, and cognitive skills?

Although research in the domain of intersensory perception of nonsocial events supports the conclusion that audiovisual relations are detected in order of increasing specificity, from detection of global amodal temporal relations, to nested amodal

relations, to modality-specific audiovisual relations, there is little research in the domain of social events investigating this developmental progression. Might information in faces and voices also be detected in order of increasing specificity? For example, do infants detect the global amodal property of temporal synchrony prior to nested amodal properties including rhythm, tempo, spectral information, and properties defined by combinations of amodal properties, such as prosody, affect, age, and gender, and are these properties detected prior to arbitrary modality-specific relations such as that between the appearance of a specific face and the specific sound of that person's voice? Future research should assess this pattern of increasing specificity in detection of face-voice relations using paradigms such as the intermodal preference and habituation methods.

Developmental Improvements in Intersensory Processing of Faces and Voices with Age. There is also evidence of developmental improvement across age in detection of audiovisual relations in faces and voices. The intermodal preference and habituation methods indicate that intersensory processing skills improve across the first 6 months of life, with the detection of amodal properties becoming more refined and detected at greater levels of detail across development. The McGurk task demonstrates improvement in audiovisual integration of phonemes with greater detection of the McGurk illusion across age and the speech in noise task indicates greater detection of audible speech in the context of noise across age. However, due to the use of cross-sectional designs and a group differences approach, the developmental trajectories for these skills is unknown. Future research should examine developmental trajectories of various intersensory processing skills, particularly using a longitudinal approach.

Developmental Impairments in Atypical Development. Finally, most paradigms have contributed either directly or indirectly to the proposal that intersensory processing serves as a foundation for language, social, and cognitive outcomes (Bahrick & Lickliter, 2012; Bahrick et al., 2020). The intermodal preference method has provided some evidence of links between detection of face-voice synchrony and concurrent language outcomes in both typically and atypically developing children. The habituation method has shown that temporal synchrony between object labels and movements facilitates learning of syllable-object relations. The McGurk task illustrates that greater audiovisual integration of incongruent audible and visible speech information is related to greater dual attention and academic readiness in school-aged children. Eye-tracking studies show that greater attention to the mouth of a speaking face is related to greater concurrent and prospective language outcomes. Finally, the individual difference approach using the MAAP and IPEP have found that early intersensory processing skills (at 6 and 12 months of age) predict later language outcomes, while controlling for well-established predictors in other domains. Future research should capitalize on advantages of individual difference approaches to establish direct links between specific intersensory processing skills and later social, cognitive, and language outcomes. Given that this research area is in its infancy, new individual difference measures are needed in order to assess detection of specific audiovisual relations. Research can address how detection of specific audiovisual relations develops across age, predicts outcomes in the domains of language, social, and cognitive functioning, and the developmental pathways through which early intersensory processing skills cascade into these more complex domains later in development.

Each paradigm also supports the conclusion that infants at risk for disabilities or children with developmental disabilities display a deficit in intersensory processing of faces and voices. A range of participants have been assessed, from infants born pre-term and at risk for autism, to children with autism, children with language disabilities, children with learning disorders, and hearing-impaired children. Thus, a large knowledge base has accrued and demonstrated that children of atypical development display a deficit in intersensory processing skills. Future research should now incorporate individual difference measures to identify infants who may be at risk for intersensory processing delays and later developing skills that rely on this foundation. To accomplish this, scientists must establish the developmental trajectories of intersensory processing of faces and voices in typically developing children. This can then serve as a basis for identifying the atypical development of intersensory processing for social events. For infants and children who fall outside the typical range of variability, interventions to improve intersensory functioning would be recommended.

Recommendations for Future Research

We recommend several future directions for intersensory processing research. First, more reliable, fine-grained measures of intersensory processing are needed. Most traditional paradigms (e.g., intermodal preference method, the habituation method) provide measures that have low reliability and are not fine-grained enough to index meaningful changes in intersensory processing across time for individual participants (Colombo et al., 2004). Such measures (e.g., gain due to visual speech in noise, eye-tracking variables, etc.) can then be used for assessing both individual-level and group-level differences. Second, using new fine-grained individual differences measures we

need to establish the typical developmental trajectories of these skills and this will allow researchers to identify children at risk for atypical development. Third, we recommend testing models of developmental pathways from basic intersensory processing skills to later developmental outcomes. Further, intersensory processing likely provides a foundation for developmental outcomes across a wide range of domains (language, cognitive skills, social functioning, executive functions, school readiness, etc.). Fourth, the role of intersensory processing in shaping development in all of these domains and more should be explored. Paradigms that assess multiple dependent measures (e.g., both speed and accuracy) are ideal and can provide a variety of indices for complementary skills that may contribute to developmental outcomes. Fifth, once models are tested and developmental pathways clarified, we can begin to develop interventions for these outcomes by targeting earlier developing skills that cascade to these outcomes. We should also focus on how intersensory processing skills can be trained. These directions all require the use of fine-grained individual difference measures.

Concluding Remarks

Decades of research on intersensory processing have provided a foundation for building theory and establishing a significant knolwedge base for the field of intersensory processing. This research was conducted with different measures and paradigms, which dictated the specific research questions and conclusions that could be drawn. Despite this, the preceding review demonstrates that these paradigms and measures converge to support a number of principles of development highlighting the importance of intersensory processing skills about relations between faces and voices during audiovisual speech for perception, learning, memory. This foundational research has

provided a substantial kmoweldge base and set the stage for asking new questions about the intersensory processing skills of individual children relative to one another, questions addressed by using an individual differences approach. Currently there are two individual difference measures designed for assessing intersensory processing skills in infants and children, and along with the development of additional individual difference measures, important new resarch questions can be addressed. These include assessing developmental trajectories of intersensory processing skills for typically and atypically developing children, predicting concurrent and future outcomes, including language, social, and cognitive skills, and deriving models assessing developmental pathways between specific intersensory processing skills and these later outcomes. The use of longitudinal studies assessing individual differences in interensory processing skills can provide insight into pathways to optimal developmental outcomes, identify children at risk for developmental delays in skills relying on intersensory processing, and guide the development of interventions for fostering optimal outcomes.

References

Adolphs, R. (2001). The neurobiology of social cognition. *Current Opinion in Neurobiology, 11,* 231–239.

Alley, T. R. (1981). Head shape and the perception of cuteness. *Developmental Psychology, 17*(5), 650–654. https://doi.org/10.1037/0012-1649.17.5.650

Baart, M., Bortfeld, H., & Vroomen, J. (2015). Phonetic matching of auditory and visual speech develops during childhood: Evidence from sine-wave speech. *Journal of Experimental Child Psychology, 129,* 157–164. https://doi.org/10.1016/j.jecp.2014.08.002

Baart, M., Vroomen, J., Shaw, K., & Bortfeld, H. (2013). Degrading phonetic information affects matching of audiovisual speech in adults, but not in infants. *Cognition, 130*(1), 31–43. https://doi.org/10.1016/j.cognition.2013.09.006

Bahrick, L. E. (1983). Infants' perception of substance and temporal synchrony in multimodal events. *Infant Behavior and Development, 6*, 429–451. https://doi.org/10.1016/S0163-6383(83)90241-2

Bahrick, L. E. (1987). Infants' intermodal perception of two levels of temporal structure in natural events. *Infant Behavior and Development, 10*(4), 387–416. https://doi.org/10.1016/0163-6383(87)90039-7

Bahrick, L. E. (1988). Intermodal learning in infancy: learning on the basis of two kinds of invariant relations in audible and visible events. *Child Development, 59*(1), 197–209. https://doi.org/10.1111/j.1467-8624.1988.tb03208.x

Bahrick, L. E. (1992). Infants' perceptual differentiation of amodal and modality-specific audio-visual relations. *Journal of Experimental Psychology, 53*, 180–199.

Bahrick, L. E. (2000). Increasing specificity in the development of intermodal perception. In D. Muir & A. Slater (Eds.), *Infant Development: The Essential Readings* (pp. 117–136). Blackwell Publishers.

Bahrick, L. E. (2001). Increasing specificity in perceptual development: Infants' detection of nested levels of multimodal stimulation. *Journal of Experimental Child Psychology, 79*, 253–270. https://doi.org/10.1006/jecp.2000.2588

Bahrick, L. E. (2002). Generalization of learning in three-and-a-half-month-old infants on the basis of amodal relations. *Child Development, 73*(3), 667–681. https://doi.org/10.1111/1467-8624.00431

Bahrick, L. E. (2004). The development of perception in a multimodal environment. In *Theories of Infant Development* (pp. 90–120). Blackwell Publishing.

Bahrick, L. E. (2010). Intermodal perception and selective attention to intersensory redundancy: Implications for typical social development and autism. In G. Bremner & T. D. Wachs (Eds.), *Blackwell handbook of infant development, 2nd ed.* (Vol. 1, pp. 120–166). Blackwell Publishing. https://doi.org/10.1002/9781444327564.ch4

Bahrick, L. E., Flom, R., & Lickliter, R. (2002). Intersensory redundancy facilitates discrimination of tempo in 3-month-old infants. *Developmental Psychobiology, 41*(4), 352–363. https://doi.org/10.1002/dev.10049

Bahrick, L. E., Hernandez-Reif, M., & Flom, R. (2005). The development of infant learning about specific face-voice relations. *Developmental Psychology, 41*(3), 541–552. https://doi.org/10.1037/0012-1649.41.3.541

Bahrick, L. E., Krogh-Jespersen, S., Argumosa, M. A., & Lopez, H. (2014). Intersensory redundancy hinders face discrimination in preschool children: Evidence for visual facilitation. *Developmental Psychology, 50*(2), 414–421. https://doi.org/10.1037/a0033476

Bahrick, L. E., & Lickliter, R. (2000). Intersensory redundancy guides attentional selectivity and perceptual learning in infancy. *Developmental Psychology, 36*(2), 190–201. https://doi.org/10.1037/0012-1649.36.2.190

Bahrick, L. E., & Lickliter, R. (2002). Intersensory redundancy guides early perceptual and cognitive development. In R. V. Kail (Ed.), *Advances in child development and behavior* (pp. 153–187). Academic Press. https://doi.org/10.1016/S0065-2407(02)80041-6

Bahrick, L. E., & Lickliter, R. (2004). Infants' perception of rhythm and tempo in unimodal and multimodal stimulation: A developmental test of the intersensory redundancy hypothesis. *Cognitive, Affective, and Behavioral Neuroscience, 4*(2), 137–147.

Bahrick, L. E., & Lickliter, R. (2012). The role of intersensory redundancy in early perceptual, cognitive, and social development. In A. Bremner, D. J. Lewkowicz, & C. Spence (Eds.), *Multisensory development* (pp. 183–205). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199586059.003.0008

Bahrick, L. E., & Lickliter, R. (2014). Learning to attend selectively: The dual role of intersensory redundancy. *Current Directions in Psychological Science, 23*(6), 414–420. https://doi.org/10.1177/0963721414549187

Bahrick, L. E., Lickliter, R., Castellanos, I., & Vaillant-Molina, M. (2010). Increasing task difficulty enhances effects of intersensory redundancy: Testing a new prediction of the Intersensory Redundancy Hypothesis. *Developmental Science, 13*(5), 731–737. https://doi.org/10.1111/j.1467-7687.2009.00928.x

Bahrick, L. E., Lickliter, R., & Flom, R. (2006). Up versus down: The role of intersensory redundancy in the development of infants' sensitivity to the orientation of moving objects. *Infancy, 9*(1), 73–96. https://doi.org/10.1207/s15327078in0901_4

Bahrick, L. E., Lickliter, R., & Todd, J. T. (2020). The development of multisensory attention skills: Individual differences, developmental outcomes, and applications. In J. J. Lockman & C. S. Tamis-LeMonda (Eds.), *The Cambridge Handbook of Infant Development* (pp. 303–338). Cambridge University Press.

Bahrick, L. E., McNew, M. E., Pruden, S. M., & Castellanos, I. (2019). Intersensory redundancy promotes infant detection of prosody in infant-directed speech. *Journal of Experimental Child Psychology, 183*, 295–309. https://doi.org/10.1016/j.jecp.2019.02.008

Bahrick, L. E., Netto, D., & Hernandez-Reif, 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

Bahrick, L. E., & Pickens, J. N. (1988). Classification of bimodal English and Spanish language passages by infants. *Infant and Child Development, 11*(3), 277–296. https://doi.org/10.1016/0163-6383(88)90014-8

Bahrick, L. E., & Pickens, J. N. (1994). Amodal relations: The basis for intermodal perception and learning. In David J. Lewkowicz & R. Lickliter (Eds.), *The development of intersensory perception: Comparative perspectives* (pp. 205–233). Lawrence Erlbaum Associates.

Bahrick, L. E., Soska, K. C., & Todd, J. T. (2018). Assessing individual differences in the speed and accuracy of intersensory processing in young children: The Intersensory Processing Efficiency Protocol. *Developmental Psychology, 54*(12), 2226–2239. https://doi.org/10.1037/dev0000575

Bahrick, L. E., & Todd, J. T. (2012). Multisensory processing in autism spectrum disorders: Intersensory processing disturbance as a basis for atypical development. In B. E. Stein (Ed.), *The new handbook of multisensory processes* (pp. 657–674). MIT Press.

Bahrick, L. E., Todd, J. T., & Soska, K. C. (2018). The Multisensory Attention Assessment Protocol (MAAP): Characterizing individual differences in multisensory attention skills in infants and children and relations with language and cognition. *Developmental Psychology*, *54*(12), 2207–2225. https://doi.org/10.1037/dev0000594

Bahrick, L. E., Walker, A. S., & Neisser, U. (1981). Selective looking by infants. *Cognitive Psychology*, *13*(3), 377–390. https://doi.org/10.1016/0010-0285(81)90014-1

Barutchu, A., Toohey, S., Shivdasani, M. N., Fifer, J. M., Crewther, S. G., Grayden, D. B., & Paolini, A. G. (2019). Multisensory perception and attention in school-age children. *Journal of Experimental Child Psychology*, *180*, 141–155. https://doi.org/10.1016/j.jecp.2018.11.021

Bebko, J. M., Weiss, J. A., Demark, J. L., & Gomez, P. (2006). Discrimination of temporal synchrony in intermodal events by children with autism and children with developmental disabilities without autism. *Journal of Child Psychology and Psychiatry and Allied Disciplines*, *47*(1), 88–98. https://doi.org/10.1111/j.1469-7610.2005.01443.x

Berdasco-Muñoz, E., Nazzi, T., & Yeung, H. H. (2019). Visual scanning of a talking face in preterm and full-term infants. *Developmental Psychology*, *55*(7), 1353–1361. https://doi.org/10.1037/dev0000737

Boliek, C., Keintz, C., Norrix, L., & Obrzut, J. (2010). Auditory-visual perception of speech in children with learning disabilities: The McGurk effect. *Canadian Journal of Speech-Language Pathology and Audiology*, *34*(2), 124–131.

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.), *Speech Reading by Humans & Machines* (NATO ASI S, 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

Caron, A. J., Caron, R. F., & Maclean, D. J. (1988). Infant discrimination of naturalistic emotional expressions: The role of face and voice. *Child Development*, *59*(3), 604–616. https://www.jstor.org/stable/1130560

Colombo, J., Shaddy, D. J., Richman, W. A., Maikranz, J. M., & Blaga, O. M. (2004). The developmental course of habituation in infancy and preschool outcome. *Infancy*, *5*(1), 1–38. https://doi.org/10.1207/s15327078in0501_1

Curtindale, L. M., Bahrick, L. E., Lickliter, R., & Colombo, J. (2019). Effects of multimodal synchrony on infant attention and heart rate during events with social and nonsocial stimuli. *Journal of Experimental Child Psychology*, *178*, 283–294. https://doi.org/10.1016/j.jecp.2018.10.006

Dawson, G., Toth, K., Abbott, R., Osterling, J., Munson, J., Estes, A., & Liaw, J. (2004). Early social attention impairments in autism: Social orienting, joint attention, and attention to distress. *Developmental Psychology*, *40*(2), 271–283. https://doi.org/10.1037/0012-1649.40.2.271

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

Dodd, B. (1979). Lip reading in infants: Attention to speech presented in- and out-of-synchrony. *Cognitive Psychology*, *11*(4), 478–484. https://doi.org/10.1016/0010-0285(79)90021-5

Dodd, B., Mcintosh, B., Erdener, D., & Burnham, D. (2008). Perception of the auditory-visual illusion in speech perception by children with phonological disorders. *Clinical Linguistics and Phonetics*, *22*(1), 69–82. https://doi.org/10.1080/02699200701660100

Dupont, S., Aubin, J., & Ménard, L. (2005). A study of the McGurk effect in 4 and 5-year-old French Canadian children. *Linguistics*, *40*, 1–17.

Edgar, E. V., Todd, J. T., & Bahrick, L. E. (n.d.-a). *Intersensory matching of faces and voices in infancy predicts language outcomes in young children*. Manuscript under review.

Edgar, E. V., Todd, J. T., & Bahrick, L. E. (n.d.-b). *Intersensory Processing of Social Events in 6-month-olds Predicts Language Outcomes at 18, 24, and 36 Months of Age*. Manuscript under review.

Erber, N. P. (1969). Interaction of audition and vision in the recognition of oral speech stimuli. *Journal of Speech and Hearing Research*, *12*(2), 423–425. https://doi.org/10.1044/jshr.1202.423

Erber, N. P. (1971). Auditory and audiovisual reception of words in low-frequency noise by children with normal hearing and by children with impaired hearing. *Journal of Speech and Hearing Research*, *14*(3), 496–512. https://doi.org/10.1044/jshr.1403.496

Feldman, J. I., Kuang, W., Conrad, J. G., Tu, A., Santapuram, P., Simon, D. M., Foss-Feig, J. H., Kwakye, L. D., Stevenson, R. A., Wallace, M. T., & Woynaroski, T. G. (2019). Brief report: Differences in multisensory integration covary with sensory responsiveness in children with and without autism spectrum disorder. *Journal of Autism and Developmental Disorders, 49*, 397–403. https://doi.org/10.1007/s10803-018-3667-x

Fernald, A., Zangl, R., Portillo, A. L., & Marchman, V. A. (2008). Looking while listening: Using eye movements to monitor spoken language. In I. A. Sekerina, E. M. Fernandez, & H. Clahsen (Eds.), *Developmental Psycholinguistics: Online methods in children's language processing* (pp. 97–135). John Benjamins.

Flom, R., & Bahrick, L. E. (2007). The development of infant discrimination of affect in multimodal and unimodal stimulation: The role of intersensory redundancy. *Developmental Psychology*, *43*(1), 238–252. https://doi.org/10.1037/0012-1649.43.1.238

Flom, R., & Bahrick, L. E. (2010). The effects of intersensory redundancy on attention and memory: Infants' long-term memory for orientation in audiovisual events. *Developmental Psychology*, *46*(2), 428–436. https://doi.org/10.1037/a0018410

Flom, R., Bahrick, L. E., & Pick, A. D. (2018). Infants discriminate the affective expressions of their peers: The roles of age and familiarization time. *Infancy*, *23*(5), 692–707. https://doi.org/10.1111/infa.12246

Flom, R., Whipple, H., & Hyde, D. (2009). Infants' intermodal perception of Ccanine (Canis familairis) facial expressions and vocalizations. *Developmental Psychology, 45*(4), 1143–1151. https://doi.org/10.1037/a0015367

Flom, R., & Whiteley, M. O. (2014). The dynamics of intermodal matching: Seven- and 12-month-olds' intermodal matching of affect. In *European Journal of Developmental Psychology* (Vol. 11, Issue 1, pp. 111–119). Taylor & Francis. https://doi.org/10.1080/17405629.2013.821059

Fort, M., Ayneto-Gimeno, A., Escrichs, A., & Sebastian-Galles, N. (2017). Impact of bilingualism on infants' ability to learn from talking and nontalking faces. *Language Learning*, *68*, 31–57. https://doi.org/10.1111/lang.12273

Gogate, L. J., & Bahrick, L. E. (1998). Intersensory redundancy facilitates learning of arbitrary relations between vowel sounds and objects in seven-month-old infants. *Journal of Experimental Child Psychology*, *69*(2), 133–149. https://doi.org/10.1006/jecp.1998.2438

Gogate, L. J., & Bahrick, L. E. (2001). Intersensory redundancy and 7-month-old infants' memory for arbitrary syllable-object relations. *Infancy*, *2*(2), 219–231. https://doi.org/10.1207/S15327078IN0202_7

Gogate, L. J., Bahrick, L. E., & Watson, J. D. (2000). A study of multimodal motherese: The role of temporal synchrony between verbal labels and gestures. *Child Development*, *71*(4), 878–894. https://doi.org/10.1111/1467-8624.00197

Gogate, L. J., & Hollich, G. (2010). Invariance detection within an interactive system: A perceptual gateway to language development. *Psychological Review*, *117*(2), 496–516. https://doi.org/10.1037/a0019049

Grant, K. W., & Seitz, P.-F. (2000). The use of visible speech cues for improving auditory detection of spoken sentences. *The Journal of the Acoustical Society of America*, *108*(3), 1197. https://doi.org/10.1121/1.1288668

Grossman, R. B., Steinhart, E., Mitchell, T., & McIlvane, W. (2015). "Look who's talking!" Gaze patterns for implicit and explicit audio-visual speech synchrony detection in children with high-functioning autism. *Autism Research*, *8*(3), 307–316. https://doi.org/10.1002/aur.1447

Guellaï, B., Streri, A., Chopin, A., Rider, D., & Kitamura, C. (2016). Newborns' sensitivity to the visual aspects of infant-directed speech: Evidence from point-line displays of talking faces. *Journal of Experimental Psychology: Human Perception and Performance*, *42*(9), 1275–1281. https://doi.org/10.1037/xhp0000208

Guiraud, J. A., Tomalski, P., Kushnerenko, E., Ribeiro, H., Davies, K., Charman, T., Elsabbagh, M., Johnson, M. H., & BASIS Team. (2012). Atypical audiovisual speech integration in infants at risk for autism. *PloS One*, *7*(5), 3–8. https://doi.org/10.1371/journal.pone.0036428

Haviland, J. M., Walker-Andrews, A. S., Huffman, L. R., Toci, L., & Alton, K. (1996). Intermodal perception of emotional expressions by children with autism. *Journal of Developmental and Physical Disabilities*, *8*(1), 77–88. https://doi.org/10.1007/BF02578441

Hayes, E. A., Tiippana, K., Nicol, T. G., Sams, M., & Kraus, N. (2003). Integration of heard and seen speech: A factor in learning disabilities in children. *Neuroscience Letters*, *351*(1), 46–50. https://doi.org/10.1016/S0304-3940(03)00971-6

Hessels, R. S., Andersson, R., Hooge, I. T. C., Nyström, M., & Kemner, C. (2015). Consequences of eye color, positioning, and head movement for eye-tracking data quality in infant research. *Infancy*, *20*(6), 601–633. https://doi.org/10.1111/infa.12093

Hessels, R. S., & Hooge, I. T. C. (2019). Eye tracking in developmental cognitive neuroscience – The good, the bad and the ugly. *Developmental Cognitive Neuroscience*, *40*, Article 100710. https://doi.org/10.1016/j.dcn.2019.100710

Hillairet de Boisferon, A., Dupierrix, E., Quinn, P. C., Lœvenbruck, H., Lewkowicz, D. J., Lee, K., & Pascalis, O. (2015). Perception of multisensory gender coherence in 6- and 9-month-old infants. *Infancy*, *20*(6), 661–674. https://doi.org/10.1111/infa.12088

Hillairet de Boisferon, A., Tift, A. H., Minar, N. J., & Lewkowicz, D. J. (2017). Selective attention to a talker's mouth in infancy: role of audiovisual temporal synchrony and linguistic experience. *Developmental Science*, *20*(3). https://doi.org/10.1111/desc.12381

Hillairet de Boisferon, A., Tift, A. H., Minar, N. J., & Lewkowicz, D. J. (2018). The redeployment of attention to the mouth of a talking face during the second year of life. *Journal of Experimental Child Psychology*, *172*, 189–200. https://doi.org/10.1016/j.jecp.2018.03.009

Hirst, R. J., Stacey, J. E., Cragg, L., Stacey, P. C., & Allen, H. A. (2018). The threshold for the McGurk effect in audio-visual noise decreases with development. *Scientific Reports*, *8*(1), 1–12. https://doi.org/10.1038/s41598-018-30798-8

Horowitz, F. D. (1974). Infant attention and discrimination: Methodological and substantive issues. *Monographs of the Society for Research in Child Development*, *39*(5/6), 1. https://doi.org/10.2307/1165968

Horowitz, F. D., Paden, L., Bhana, K., & Self, P. (1972). An infant-control procedure for studying infant visual fixations. *Developmental Psychology*, *7*(1), 90. https://doi.org/10.1037/h0032855

Hyde, D. C., Jones, B. L., Flom, R., & Porter, C. L. (2011). Neural signatures of face-voice synchrony in 5-month-old human infants. *Developmental Psychobiology, 53*(4), 359–370. https://doi.org/10.1002/dev.20525

Iarocci, G., Rombough, A., Yager, J., Weeks, D. J., & Chua, R. (2010). Visual influences on speech perception in children with autism. *Autism, 14*(4), 305–320. https://doi.org/10.1177/1362361309353615

Imafuku, M., Kawai, M., Niwa, F., Shinya, Y., & Myowa, M. (2019). Audiovisual speech perception and language acquisition in preterm infants: A longitudinal study. *Early Human Development, 128*(February), 93–100. https://doi.org/10.1016/j.earlhumdev.2018.11.001

Imafuku, M., & Myowa, M. (2016). Developmental change in sensitivity to audiovisual speech congruency and its relation to language in infants. *Psychologia, 59*(4), 163–172. https://doi.org/10.2117/psysoc.2016.163

Irwin, J. R., Tornatore, L. A., Brancazio, L., & Whalen, D. H. (2011). Can children with autism spectrum disorders "hear" a speaking face? *Child Development, 82*(5), 1397–1403. https://doi.org/10.1111/j.1467-8624.2011.01619.x

Kahana-Kalman, R., & Goldman, S. (2007). Intermodal matching of emotional expressions in young children with autism. *Research in Autism Spectrum Disorders, 2*(2), 301–310. https://doi.org/10.1016/j.rasd.2007.07.004

Kahana-Kalman, R., & Walker-Andrews, A. S. (2001). The role of person familiarity in young infants' perception of emotional expressions. *Child Development, 72*(2), 352–369. https://doi.org/10.1111/1467-8624.00283

Kitamura, C., Guellaï, B., & Kim, J. (2014). Motherese by eye and ear: Infants perceive visual prosody in point-line displays of talking heads. *PLoS ONE, 9*(10). https://doi.org/10.1371/journal.pone.0111467

Knowland, V. C. P., Evans, S., Snell, C., & Rosen, S. (2016). Visual speech perception in children with language learning impairments. *Journal of Speech, Language, and Hearing Research, 59*, 1–14. https://doi.org/10.1044/2015_JSLHR-S-14-0269

Kubicek, C., De Boisferon, A. H., Dupierrix, E., Pascalis, O., Loevenbruck, H., Gervain, J., & Schwarzer, G. (2014). Cross-modal matching of audio-visual german and french fluent speech in infancy. *PLoS ONE, 9*(2). https://doi.org/10.1371/journal.pone.0089275

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

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 of the United States of America, 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

Lasky, R. E., Klein, R. E., & Martínez, S. (1974). Age and sex discriminations in five- and six-month-old infants. *Journal of Psychology: Interdisciplinary and Applied*, *88*(2), 317–324. https://doi.org/10.1080/00223980.1974.9915743

Lewkowicz, D. J. (1992). Infants' response to temporally based intersensory equivalence: The effect of synchronous sounds on visual preferences for moving stimuli. *Infant Behavior and Development*, *15*(3), 297–324. https://doi.org/10.1016/0163-6383(92)80002-C

Lewkowicz, D. J. (2000). Infants' perception of the audible, visible, and bimodal attributes of multimodal syllables. *Child Development*, *71*(5), 1241–1257. https://doi.org/10.1111/1467-8624.00226

Lewkowicz, D. J. (2003). Learning and discrimination of audiovisual events in human infants: The hierarchical relation between intersensory temporal synchrony and rhythmic pattern cues. *Developmental Psychology*, *39*(5), 795–804. https://doi.org/10.1037/0012-1649.39.5.795

Lewkowicz, D. J. (2010). Infant perception of audio-visual speech synchrony. *Developmental Psychology*, *46*(1), 66–77. https://doi.org/10.1037/a0015579

Lewkowicz, D. J., & Ghazanfar, A. A. (2006). The decline of cross-species intersensory perception in human infants. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(17), 6771–6774. https://doi.org/10.1073/pnas.0602027103

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 of the United States of America*, *109*(5), 1431–1436. https://doi.org/10.1073/pnas.1114783109

Lewkowicz, D. J., Leo, I., & Simion, F. (2010). Intersensory perception at birth: Newborns match nonhuman primate faces and voices. *Infancy*, *15*(1), 46–60. https://doi.org/10.1111/j.1532-7078.2009.00005.x

Lewkowicz, D. J., Minar, N. J., Tift, A. H., & Brandon, M. (2015). Perception of the multisensory coherence of fluent audiovisual speech in infancy: Its emergence and the role of experience. *Journal of Experimental Child Psychology*, *130*, 147–162. https://doi.org/10.1016/j.jecp.2014.10.006

Lewkowicz, D. J., & Pons, F. (2013). Recognition of amodal language identity emerges in infancy. *International Journal of Behavioral Development*, *37*(2), 90–94. https://doi.org/10.1177/0165025412467582

Lewkowicz, David J. (2000). The Development of Intersensory Temporal Perception: An Epigenetic Systems/Limitations View. *Psychological Bulletin*, *126*(2), 281–308. https://doi.org/10.1037/0033-2909.126.2.281

Lewkowicz, David J., & Lickliter, R. (1994). *The Development of Intersensory Perception: Comparative Perspectives* (1st ed.). Psychology Press.

Loveland, K. A., Tunali-Kotoski, B., Chen, R., Brelsford, K. A., Ortegon, J., & Pearson, D. A. (1995). Intermodal perception of affect in persons with autism or Down syndrome. *Development and Psychopathology*, *7*(3), 409–418. https://doi.org/10.1017/S095457940000660X

Macdonald, J., & McGurk, H. (1978). Visual influences on speech perception processes. *Perception & Psychophysics*, *24*(3), 253–257. https://doi.org/10.3758/BF03206096

Macleod, A., & Summerfield, Q. (1987). Quantifying the contribution of vision to speech perception in noise. *British Journal of Audiology*, *21*(2), 131–141. https://doi.org/10.3109/03005368709077786

Mason, G. M., Goldstein, M. H., & Schwade, J. A. (2019). The role of multisensory development in early language learning. *Journal of Experimental Child Psychology*, *183*, 48–64. https://doi.org/10.1016/j.jecp.2018.12.011

Massaro, D. W. (1984). Children's perception of visual and auditory speech. *Child Development*, *55*(5), 1777–1788. https://doi.org/10.2307/1129925

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, *264*(5588), 746–748. https://www-nature-com.libproxy1.usc.edu/articles/264746a0.pdf%0Ahttps://www.nature.com/articles/264746a0.pdf

Mongillo, E. A., Irwin, J. R., Whalen, D. H., Klaiman, C., Carter, A. S., & Schultz, R. T. (2008). Audiovisual processing in children with and without autism spectrum disorders. *Journal of Autism and Developmental Disorders*, *38*(7), 1349–1358. https://doi.org/10.1007/s10803-007-0521-y

Morin-Lessard, E., Poulin-Dubois, D., Segalowitz, N., & Byers-Heinlein, K. (2019). Selective attention to the mouth of talking faces in monolinguals and bilinguals aged 5 months to 5 years. *Developmental Psychology*, *55*(8), 1640–1655. https://doi.org/10.1037/dev0000750

Nath, A. R., & Beauchamp, M. S. (2011). Dynamic changes in superior temporal sulcus connectivity during perception of noisy audiovisual speech. *Journal of Neuroscience*, *31*(5), 1704–1714. https://doi.org/10.1523/JNEUROSCI.4853-10.2011

Nath, A. R., Fava, E. E., & Beauchamp, M. S. (2011). Neural correlates of interindividual differences in children's audiovisual speech perception. *Journal of Neuroscience*, *31*(39), 13963–13971. https://doi.org/10.1523/JNEUROSCI.2605-11.2011

Norrix, L. W., Plante, E., Vance, R., & Boliek, C. A. (2007). Auditory-visual integration for speech by children with and without specific language impairment. *Journal of Speech, Language, and Hearing Research*, *50*(6), 1639–1651. https://doi.org/10.1044/1092-4388(2007/111)

O'Neill, J. J. (1954). Contributions of the visual components of oral symbols to speech comprehension. *The Journal of Speech and Hearing Disorders*, *19*(4), 429–439. https://doi.org/10.1044/jshd.1904.429

Oakes, L. M. (2010). Infancy guidelines for publishing eye-tracking data. *Infancy*, *15*(1), 1–5. https://doi.org/10.1111/j.1532-7078.2010.00030.x

Oakes, L. M. (2012). Advances in eye tracking in infancy research. *Infancy*, *17*(1), 1–8. https://doi.org/10.1111/j.1532-7078.2011.00101.x

Oakes, L. M., & Rakison, D. H. (2020). *Developmental cascades: Building the infant mind*. Oxford University Press.

Patten, E., Labban, J. D., Casenhiser, D. M., & Cotton, C. L. (2016). Synchrony Detection of Linguistic Stimuli in the Presence of Faces: Neuropsychological Implications for Language Development in ASD. *Developmental Neuropsychology*, *41*(5–8), 362–374. https://doi.org/10.1080/87565641.2016.1243113

Patten, E., Watson, L. R., & Baranek, G. T. (2014). Temporal synchrony detection and associations with language in young children with ASD. *Autism Research and Treatment*, 678346. https://doi.org/10.1155/2014/678346

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. (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

Peterson, G. E., & Barney, H. L. (1952). Control methods used in a study of vowels. *The Journal of the Acoustical Society of America*, *24*(2), 175–184. https://doi.org/10.1121/1.1906875

Pickens, J., Field, T., Nawrocki, T., Martinez, A., Soutullo, D., & Gonzalez, J. (1994). Full-term and preterm infants' perception of face-voice synchrony. *Infant Behavior and Development*, *17*(4), 447–455. https://doi.org/10.1016/0163-6383(94)90036-1

Pons, F., Andreu, L., Sanz-Torrent, M., Buil-Legaz, L., & Lewkowicz, D. J. (2013). Perception of audio-visual speech synchrony in Spanish-speaking children with and without specific language impairment. *Journal of Child Language*, *40*(3), 687–700. https://doi.org/10.1017/S0305000912000189

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

Pons, F., Bosch, L., & Lewkowicz, D. J. (2019). Twelve-month-old infants' attention to the eyes of a talking face is associated with communication and social skills. *Infant Behavior and Development*, *54*(December), 80–84. https://doi.org/10.1016/j.infbeh.2018.12.003

Pons, F., & Lewkowicz, D. J. (2014). Infant perception of audio-visual speech synchrony in familiar and unfamiliar fluent speech. *Acta Psychologica*, *149*, 142–147. https://doi.org/10.1016/j.actpsy.2013.12.013

Pons, F., Lewkowicz, D. J., Soto-Faraco, S., & Sebastián-Gallés, N. (2009). Narrowing of intersensory speech perception in infancy. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(26), 10598–10602. https://doi.org/10.1073/pnas.0904134106

Pons, F., Sanz-Torrent, M., Ferinu, L., Birulés, J., & Andreu, L. (2018). Children with SLI can exhibit reduced attention to a talker's mouth. *Language Learning*, *68*, 180–192. https://doi.org/10.1111/lang.12276

Poulin-Dubois, D., Serbin, L. A., & Derbyshire, A. (1998). Toddlers ' intermodal and verbal knowledge about gender. *Merrill-Palmer Quarterly*, *44*(3), 338–354. http://www.jstor.com/stable/23093706

Poulin-Dubois, D., Serbin, L. A., Kenyon, B., & Derbyshire, A. (1994). Infants' intermodal knowledge about gender. *Developmental Psychology*, *30*(3), 436–442. https://doi.org/10.1037/0012-1649.30.3.436

Reynolds, G. D., Bahrick, L. E., Lickliter, R., & Guy, M. W. (2014). Neural correlates of intersensory processing in 5-month-old infants. *Developmental Psychobiology*, *56*(3), 355–372. https://doi.org/10.1002/dev.21104

Richoz, A. R., Quinn, P. C., De Boisferon, A. H., Berger, C., Loevenbruck, H., Lewkowicz, D. J., Lee, K., Dole, M., Caldara, R., & Pascalis, O. (2017). Audio-visual perception of gender by infants emerges earlier for adult-directed speech. *PLoS ONE*, *12*(1), 1–15. https://doi.org/10.1371/journal.pone.0169325

Righi, G., Tenenbaum, E. J., McCormick, C., Blossom, M., Amso, D., & Sheinkopf, S. J. (2018). Sensitivity to audio-visual synchrony and its relation to language abilities in children with and without ASD. *Autism Research*, *11*(4), 645–653. https://doi.org/10.1002/aur.1918

Rose, S. A., Feldman, J. F., Jankowski, J. J., & Van Rossem, R. (2012). Information processing from infancy to 11years: Continuities and prediction of IQ. *Intelligence*, *40*(5), 445–457. https://doi.org/10.1016/j.intell.2012.05.007

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

Ross, L. A., Molholm, S., Blanco, D., Gomez-Ramirez, M., Saint-Amour, D., & Foxe, J. J. (2011). The development of multisensory speech perception continues into the late childhood years. *European Journal of Neuroscience*, *33*(12), 2329–2337. https://doi.org/10.1111/j.1460-9568.2011.07685.x

Sekiyama, K., & Burnham, D. (2008). Paper: 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

Shaw, K., Baart, M., Depowski, N., & Bortfeld, H. (2015). Infants' preference for native audiovisual speech dissociated from congruency preference. *PLoS ONE*, *10*(4), 1–11. https://doi.org/10.1371/journal.pone.0126059

Shic, F., Macari, S., & Chawarska, K. (2014). Speech disturbs face scanning in 6-month-old infants who develop autism spectrum disorder. *Biological Psychiatry*, *75*(3), 231–237. https://doi.org/10.1016/j.biopsych.2013.07.009

Slater, A., Quinn, P. C., Brown, E., & Hayes, R. (1999). Intermodal perception at birth: Intersensory redundancy guides newborn infants' learning of arbitrary auditory-visual pairings. *Developmental Science*, *2*(3), 333–338. https://doi.org/10.1111/1467-7687.00079

Soken, N. H., & Pick, A. D. (1992). Intermodal perception of happy and angry expressive behaviors by seven-month-old infants. *Child Development*, *63*(4), 787–795. https://doi.org/10.1111/j.1467-8624.1992.tb01661.x

Spelke, E. (1976). Infants' intermodal perception of events. *Cognitive Psychology*, *8*, 553–560. https://doi.org/10.1016/0010-0285(76)90018-9

Spelke, E. S., & Cortelyou, A. (1981). Perceptual aspects of social knowing: Looking and listening in infancy. In M. E. Lamb & L. R. Sherrod (Eds.), *Infant social cognition: Empirical and theoretical considerations* (pp. 60–83). Erlbaum.

Spelke, E. S., Smith Born, W., & Chu, F. (1983). Perception of moving, sounding objects by four-month-old infants. *Perception*, *12*(6), 719–732. https://doi.org/10.1068/p120719

Spelke, E. S. (1979). Perceiving bimodally specified events in infancy. *Developmental Psychology*, *15*(6), 626–636. https://doi.org/10.1037/0012-1649.15.6.626

Spelke, E. S. (1981). The infant's acquisition of knowledge of bimodally specified events. *Journal of Experimental Child Psychology*, *31*(2), 279–299. https://doi.org/10.1016/0022-0965(81)90018-7

Spelke, E. S., & Owsley, C. J. (1979). Intermodal exploration and knowledge in infancy. *Infant Behavior and Development*, *2*(1), 13–27. https://doi.org/10.1016/S0163-6383(79)80004-1

Stevenson, R. A., Baum, S. H., Segers, M., Ferber, S., Barense, M. D., & Wallace, M. T. (2017). Multisensory speech perception in autism spectrum disorder: From phoneme to whole-word perception. *Autism Research*, *10*(7), 1280–1290. https://doi.org/10.1002/aur.1776

Stevenson, R. A., Siemann, J. K., Woynaroski, T. G., Schneider, B. C., Eberly, H. E., Camarata, S. M., & Wallace, M. T. (2014). Arrested development of audiovisual speech perception in autism spectrum disorders. *Journal of Attention Disorders*, *44*(6), 1470–1477. https://doi.org/10.1007/s10803-013-1992-7

Sumby, W. H., & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *Journal of the Acoustical Society of America*, *26*(2), 212–215. https://doi.org/10.1121/1.1907309

Taylor, N., Isaac, C., & Milne, E. (2010). A comparison of the development of audiovisual integration in children with autism spectrum disorders and typically developing children. *Journal of Autism and Developmental Disorders*, *40*(11), 1403–1411. https://doi.org/10.1007/s10803-010-1000-4

Tenenbaum, E. J., Shah, R. J., Sobel, D. M., Malle, B. F., & Morgan, J. L. (2013). Increased focus on the mouth among infants in the first year of life: A longitudinal eye-tracking study. *Infancy*, *18*(4), 534–553. https://doi.org/10.1111/j.1532-7078.2012.00135.x

Tenenbaum, E. J., Sobel, D. M., Sheinkopf, S. J., Malle, B. F., & Morgan, J. L. (2015). Attention to the mouth and gaze following in infancy predict language development. *Journal of Child Language*, *42*(6), 1173–1190. https://doi.org/10.1017/S0305000914000725

Todd, J. T., & Bahrick, L. E. (2020). Characterizing multisensory attention in early development: Individual differences, trajectories, and relations with outcomes. *International Congress on Infant Studies*, Virtual.

Todd, J. T., McNew, M. E., Soska, K. C., & Bahrick, L. E. (2017). Assessing the cost of competing stimulation on attention to multimodal events: Longitudinal findings from 3 to 12 months. *Society for Research in Child Development*, Austin, TX.

Tomalski, P., Ribeiro, H., Ballieux, H., Axelsson, E. L., Murphy, E., Moore, D. G., & Kushnerenko, E. (2013). Exploring early developmental changes in face scanning patterns during the perception of audiovisual mismatch of speech cues. *European Journal of Developmental Psychology*, *10*(5), 611–624. https://doi.org/10.1080/17405629.2012.728076

Tremblay, C., Champoux, F., Voss, P., Bacon, B. A., Lepore, F., & Théoret, H. (2007). Speech and non-speech audio-visual illusions: A developmental study. *PLoS ONE*, *2*(8). https://doi.org/10.1371/journal.pone.0000742

Tsang, T., Atagi, N., & Johnson, S. P. (2018). Selective attention to the mouth is associated with expressive language skills in monolingual and bilingual infants. *Journal of Experimental Child Psychology*, *169*, 93–109. https://doi.org/10.1016/j.jecp.2018.01.002

Vaillant-Molina, M., & Bahrick, L. E. (2012). The role of intersensory redundancy in the emergence of social referencing in 5 1/2-month-old infants. *Developmental Psychology*, *48*(1), 1–9. https://doi.org/10.1037/a0025263

Vaillant-Molina, M., Bahrick, L. E., & Flom, R. (2013). Young infants match facial and vocal emotional expressions of other infants. *Infancy*, *18*, 97–111. https://doi.org/10.1111/infa.12017

Walker-Andrews, A. S., & Grolnick, W. (1983). Discrimination of vocal expressions by young infants*. *Infant Behavior and Development*, *6*(4), 491–498. https://doi.org/10.1016/S0163-6383(83)90331-4

Walker-Andrews, A. S., & Lennon, E. (1991). Infants' discrimination of vocal expressions: Contributions of auditory and visual information. *Infant Behavior and Development*, *14*(2), 131–142. https://doi.org/10.1016/0163-6383(91)90001-9

Walker-Andrews, A. S. (1986). Intermodal perception of expressive behaviors: Relation of eye and voice? *Developmental Psychology*, *22*(3), 373–377. https://doi.org/10.1037/0012-1649.22.3.373

Walker-Andrews, A. S. (1997). Infants' perception of expressive behaviors: Differentiation of multimodal information. *Psychological Bulletin*, *121*(3), 437–456. https://doi.org/10.1037/0033-2909.121.3.437

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

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

Werker, J. F. (2018). Perceptual beginnings to language acquisition. *Applied Psycholinguistics*, *39*(4), 703–728. https://doi.org/10.1017/S0142716418000152

Wilcox, T., Stubbs, J. A., Wheeler, L., & Alexander, G. M. (2013). Infants' scanning of dynamic faces during the first year. *Infant Behavior and Development*, *36*(4), 513–516. https://doi.org/10.1016/j.infbeh.2013.05.001

Williams, J. H. G., Massaro, D. W., Peel, N. J., Bosseler, A., & Suddendorf, T. (2004). Visual-auditory integration during speech imitation in autism. *Research in Developmental Disabilities*, *25*(6), 559–575. https://doi.org/10.1016/j.ridd.2004.01.008

Woynaroski, T. G., Kwakye, L. D., Foss-Feig, J. H., Stevenson, R. A., Stone, W. L., & Wallace, M. T. (2013). Multisensory speech perception in children with autism spectrum disorders. *Journal of Autism and Developmental Disorders*, *43*(12), 2891–2902. https://doi.org/10.1007/s10803-013-1836-5

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

Tables

Table 1

Summary of studies and findings for intersensory processing of faces and voices as assessed by the intermodal preference method.

Reference	Participant/ Age	Stimuli	Method	Event Properties Assessed	DV	Results	Conclusions
Detection of global amodal properties							
Spelke, 1976	3- to 4-month-olds	Two identical women side-by-side each playing peek-a-boo	2-screen intermodal preference method	Audiovisual (AV) synchrony	Proportion of total looking time (PTLT) to sound-synchronous film	PTLT to sound-synchronous film was significantly greater than chance	Infants perceive invariant relations (temporal synchrony) between auditory and visual information
Dodd, 1979	2.5- to 4-month-olds	A visual display of live experimenter speaking along with soundtrack shifting in and out of synchrony every 60-s	Live experimenter in soundproof booth. Soundtrack was delayed or in-synchrony	AV synchrony	Percent of attention to speech both in- and out-of-synch	Infants attended significantly less to the speech presented out-of-synchrony than to the speech	Infants spend a greater amount of time attending to synchronous speech than asynchronous speech

						presented synchronously	
Kuhl & Meltzoff, 1982	4.5- to 5-month-olds	Two identical women side-by-side each articulating a different vowel (a vs. i) in synchrony with each other or two tones (spectral information removed). Both events played in-synchrony with a soundtrack matching one of the vowels or matching one of the tones.	2-screen intermodal preference method	1. Voice: AV spectral information (synchrony controlled) 2. Tone: AV amplitude, duration (synchrony controlled)	Percent of total fixation time to the mouth movements matching the vowel sound	Infants fixated the mouth movements matching the vocal information significantly more than chance for voice but not tone	Infants can match mouth movements and vowel sounds on the basis of spectral information when synchrony is controlled
Patterson & Werker, 1999	4.5- to 5-month-olds	Two identical women (or two identical men) side-by-side each articulating a different vowel (a vs. i)	2-screen intermodal preference method after familiarizatio n	AV spectral information (synchrony controlled)	Percent of total looking time to the mouth movements matching the vowel sound	Infants fixated the mouth movements matching the vocal information significantly	Infants can match female and male mouth movements and vowel sounds on the basis of

		in synchrony with each other. Both events played in-synchrony with a soundtrack matching one of the vowels.				more than chance for both the woman and man	spectral information when synchrony is controlled.
Patterson & Werker, 2003	2- to 3-month-olds	Two identical women (or two identical men) side-by-side each articulating a different vowel (a vs. i) in synchrony with each other. Both events played in-synchrony with a soundtrack matching one of the vowels.	2-screen intermodal preference method after familiarizatio n	AV spectral information (synchrony controlled)	Percent of total looking time to the mouth movements matching the vowel sound	Infants fixated the mouth movements matching the vocal information significantly more than chance for both the woman and man	Infants can match female and male mouth movements and vowel sounds on the basis of spectral information when synchrony is controlled
Lewkowicz et al., 2010	1- to 3-day-olds	Two identical rhesus monkeys side-by-side each articulating a different sound (coo vs.	2-screen intermodal preference method	1. Coo/grunt: AV spectral information, duration, and event offset	Proportion of looking to the mouth movements matching the sound (coo	Newborn infants looked to the monkey face matching the vocal information	Newborn infants show cross-species matching of mouth movements and

		grunt). Both events started in-synchrony with the soundtrack, but the offset matched only one of the sounds or one of the complex tones (spectral information removed).		2. Tone: AV duration and event offset	vs. grunt) or the tone	significantly more than chance for vocalizations and complex tones	sounds on the basis of synchrony (duration, event offset) both with and without spectral information
Baart et al., 2013	5- to 15-month-olds	Two identical women side-by-side each speaking a different pseudo word in natural or sine wave speech (SWS). Both events started in-synchrony with the soundtrack, but the offset was synchronous	2-screen intermodal preference method	1. Natural speech: AV phonetic information and temporal duration/offset 2. SWS: AV temporal duration/offset	Proportion of time looking to the mouth movements matching the word	Infants looked to the mouth movements matching the vocal information significantly more than chance for natural and SWS.	Infants can match mouth movements and words in natural and SWS on the basis of AV phonetic and temporal information (in contrast, adults used phonetic information for matching more than temporal information)

		with only one of the words.					
Kitamura et al., 2014	8-month-olds	Two point-line displays of identical women side-by-side each speaking a different sentence in synchrony with each other. Both events played in synchrony with a soundtrack matching the number of syllables in one of the sentences. Altered auditory stimuli (natural vs. intonation-only/low-pass filtered speech) or visual stimuli (rigid vs. non-rigid facial movement) to remove	2-screen intermodal preference method	Auditory: 1a. Natural speech: AV prosodic information 1b. Intonation-only speech: auditory prosodic information Visual 2a. Rigid facial movement: intonation, word stress and phrasal rhythm conveyed by head motion (up-down, side to side, front to back) and voice (pitch)	Fixation time to the face with movements matching the sentence	Infants looked to the facial movements matching the vocal information significantly more than chance for both auditory stimuli, and for one visual stimulus (rigid facial movements)	Infants can match facial movements and sentences on the basis of global AV prosodic information when phonetic information is removed

		phonetic information		2b. Non-rigid facial movement: AV temporal information from internal face features (mouth/jaw, eyes, cheeks)			
Baart et al., 2015	4- to 11-year-olds	Two identical women side-by-side each speaking a different pseudo word in sine wave speech (SWS). Both events started in-synchrony with the soundtrack, but the offset matched only one of the words.	2-screen intermodal preference method. Non-speech (with SWS) followed by speech training mode (paired SWS words with natural words to inform children of phonetic information in SWS).	1. Non-speech: AV temporal offset 2. Speech training: AV temporal offset, phonetic information	Proportion of correct verbal responses indicating which mouth movements matched the word heard	The proportion of correct responses for the mouth movement matching the vocal information was significantly greater for speech training mode than for SWS at 7–9 and 9–11 years, but not 4–7 years old	Older children match mouth movements and words on the basis of phonetic information, whereas younger children match on the basis of temporal information
Guellaï et al., 2016	2-day-olds	Point-line displays of two identical	2-screen intermodal	1. Rigid facial movement: intonation,	PTLT to the facial movements	Newborns looked to the facial	Newborns match mouth movements and

		familiar women (infants' mother) side-by-side, each speaking a different sentence in synchrony with each other. Both events were synchronized with a soundtrack matching the number of syllables in one of the sentences. Altered visual stimuli (rigid or non-rigid facial movements) to remove phonetic information	preference method	word stress and phrasal rhythm conveyed by head motion and voice (controlling for temporal information from mouth/ jaw) 2.Non-rigid facial movement: AV temporal information from mouth/jaw (controlling for head motion)	matching the sentence	movements matching the vocal information significantly more than chance for rigid and non-rigid facial movements	sentences on the basis auditory and visual prosodic information (rhythm and intonation) and temporal synchrony between seen and heard mouth/jaw movements
Detection of attributes defined by a combination of amodal properties							
Lasky et al., 1974	5- to 7-month-olds	Achromatic pictures of men, women, and boys presented	2-screen intermodal preference method	Gender, age	Fixation time to the face matching the	Infants looked to the face matching the vocal	Infants can match faces and voices on the basis of

		side-by-side (man v woman, man v. boy, woman v. boy). Both pictures presented with a soundtrack matching the gender in one of the pictures			vocal information	information significantly more than chance for the pictures of the woman paired with the boy	gender. Infants can match on the basis of age but only when presented with different genders
Spelke & Owsley, 1979	3.5- to 7.5-month-olds	Live visual display of both parents side-by-side or the mother and an unfamiliar woman side-by-side. Events played with soundtrack or live-voice episode matching the voice of one of the people	Live intermodal preference method	Parental familiarity	Duration of looking to the person matching the voice	Infants looked to the parent matching the vocal information significantly more than chance, but only when both parents were paired together	Infants can match faces and voices on the basis of parental familiarity when presented with both parents
Walker, 1982	5- and 7-month-olds	Two identical women side-by-side each speaking and	2-screen intermodal preference method	Affect	PTLT to the facial movements	Infants 5 and 7 months looked to the facial movements	Infants 5 months can match faces and voices in

		gesturing in a different emotion (happy, sad, neutral, angry) in normal orientation or with faces inverted. Both events played in-synchrony or out-of-synchrony with a soundtrack matching one of the emotions.			matching the emotion	matching the vocal information significantly more than chance for synchronous, upright displays, whereas on 7-month infants also did for asynchronous but not inverted displays	presence of temporal synchrony, whereas infant 7 months can match basis of affect (removing temporal information), but not when faces are inverted (removing configuration information)
Walker-Andrews, 1986	5- and 7-month-olds	The upper 1/3 of the faces of two identical women side-by-side each speaking in a different emotion (happy vs. angry) in synchrony with each other. Both events played in-synchrony	Intermodal preference method	Affect	PTLT to the upper one-third of facial movements matching the emotion	Infants 7 (but not 5) months looked to the upper one-third of facial movements matching the vocal information significantly more than chance.	By 7 months, infants can match faces and voices on the basis of affect when the lower two-thirds of the face is occluded.

		with a soundtrack matching one of the emotions.					
Walker-Andrews et al., 1991	3.5- to 6.5-month-olds	A man and a woman standing side-by-side speaking the same nursery rhyme in synchrony with each other. Both events played in synchrony with a soundtrack matching one of the people.	Intermodal preference method	Gender	PTLT to the person matching the soundtrack	Infants 6–6.5 (but not 3.5–4) months looked to the person matching the vocal information significantly more than chance	By 6 months, infants can match faces and voices on the basis of gender
Soken & Pick, 1992	7-month-olds	Two different women side-by-side each speaking and gesturing in a different emotion (happy v. angry) in normal or point-light display. Both events played in- or	2-screen intermodal preference method	Affect	PTLT to the facial movements matching the emotion	Infants looked to the facial movements matching the vocal information significantly more than chance for synchronous, asynchronous, normal, and	Infants can match faces and voices on the basis of affect

		out-of-synchrony with a soundtrack matching one of the emotions.				point light displays	
Poulin-Dubois et al., 1994	9- and 12-month-olds	Achromatic pictures of a man and a woman presented side-by-side. Both pictures presented with a soundtrack matching one of the pictures.	2-screen intermodal preference method	Gender	Average looking time to the picture matching the vocal information	Infants 12, but not 9, months looked to the face matching the vocal information significantly more than chance	By 12 months, infants can match faces and voices on the basis of gender
Poulin-Dubois et al., 1998	18- and 24-month-olds	Achromatic pictures of adult and child male and female faces presented side-by-side (man v. boy, man v. girl, woman v. boy, woman v. girl, man v. woman, boy v. girl). Both pictures	2-screen intermodal preference method	Gender	Average looking time to the picture matching the voice	Infants looked to the face matching the vocal information significantly more than chance for adult, but not child, faces	Toddlers match faces and voices on the basis of gender for adult, but not child faces

		presented with a soundtrack matching one of the pictures					
Bahrick et al., 1998	4- and 7-month-olds	Adult and child male and female faces side-by-side (man v. boy, man v. girl, woman v. boy, woman v. girl, man v. woman, boy v. girl), each reciting the same nursery rhyme in synchrony with each other. Both events presented in normal orientation or with faces inverted, and played in synchrony with a soundtrack matching one of the people.	2-screen intermodal preference method	Age	PTLT to the person matching the voice	Infants looked to the person matching the vocal information significantly more than chance for adult and child faces, but not when they were inverted	Infants can match faces and voices that differentiate adults and children on the basis of configurational information

Kahana-Kalman & Walker-Andrews, 2001	3.5- to 4.5-month-olds	An identical woman (unfamiliar or infant's mother) side-by-side, each speaking and gesturing in a different emotion (happy v. sad). Both events played in-synchrony or out-of-synchrony with a soundtrack matching one of the emotions.	2-screen intermodal preference method	Affect	PTLT to the facial information matching the emotion	Infants looked to the facial information matching the emotion significantly more than chance for mother in- and out-of-synchrony, but not unfamiliar woman	Infants can match faces and voices of their own mother on the basis of affect
Vaillant-Molina et al., 2013	3.5- and 5-month-olds	7.5- to 8-month-old infants side-by-side each expressing a different emotion (happy/joy v. frustration/anger). Both events played in-synchrony with a soundtrack	2-screen intermodal preference method	Affect	PTLT to the facial information matching the emotion	Infants 5, but not 3.5, months looked to the facial information matching the emotion significantly more than chance	By 5 months, infants can match the faces and voices of other infants on the basis of affect

		matching one of the emotions.					
Flom & Whiteley, 2014	7- and 12-month-olds	Pictures of two different women side-by-side, each conveying a different emotion (happy v. sad). Both pictures presented with a soundtrack matching one of the emotions.	2-screen intermodal preference method.	Affect	PTLT to the facial information matching the emotion	Infants 7, but not 12, months looked to the facial information matching the emotion significantly more than chance	Infants 7 months match faces and emotions on the basis of affect, but do not at 12 months
Hillairet de Boisferon et al., 2015	6- and 9-month-olds	A woman and a man side-by-side each singing the same nursery rhyme in synchrony. Both events played in-synchrony with a soundtrack matching one the people.	2-screen intermodal preference	Gender	PTLT to the facial information matching the vocal information	Infants 9, but not 6, months looked to the facial information matching the vocal information significantly more than chance for female, but not male.	Infants 9 months match faces and voices on the basis of gender, but do not do so at 6 months

Richoz et al., 2017	6-, 9-, and 12-month-olds	A woman and a man side-by-side each singing the same nursery rhyme synchronously in adult- or infant-directed speech. Both events played in-synchrony with a soundtrack matching one the people.	2-screen intermodal preference method	Gender. Tested two possibilities of infant-directed speech: 1. facilitate matching on basis of gender or 2. draws attention to prosodic features of linguistic content	PTLT to the facial information matching the vocal information	Infants 6 months looked to the facial information matching the vocal information significantly more than chance for adult-, but not infant-directed speech, but infants 9 and 12 months did not differ depending on speech condition	Adult-directed speech facilitates face-voice matching on the basis of gender at 6 months, whereas infants 9 and 12 months match on the basis of gender for both adult- and infant-directed speech
Face-voice matching in pre-term infants and children displaying developmental disabilities							
Pickens et al., 1994	3-, 5-, and 7-month-olds born full- and pre-term	Two different women side-by-side each reciting a different song (Jingle Bells vs. Brother John) in synchrony with each other. Both	2-screen intermodal preference method	AV temporal information	PTLT to the mouth movements matching the vocal information	Full-term infants looked to the mouth movements matching the vocal information significantly more than	Full-term infants detect AV temporal synchrony across face and voice by 3 months, showing a U-shaped

		events played in-synchrony with a soundtrack matching one the songs.				chance at 3 and 7, but not 5 months, whereas pre-term infants did not at any age.	developmental pattern, but pre-term infants do not.
Loveland et al., 1995	66- to 316-month-olds with autism or Down syndrome	Two identical women side-by-side each speaking with a different emotion (happy, sad, angry, surprised, neutral). Both events played in- and out-of-synchrony with a soundtrack matching one of the emotions.	2-screen intermodal preference method	1. Synchrony: AV temporal information 2. Asynchrony: Affect	Verbal or nonverbal (point) response to indicate which woman's facial movements matched the vocal information	Children did not respond which woman's facial movements matched the vocal information significantly more than chance without synchrony. Children with autism gave fewer correct responses than children with Down syndrome for synchrony condition. All children	All children do not match facial and vocal information on the basis of affect, but do so on the basis of temporal synchrony. Children with autism are poorer at matching faces and voices than children with Down syndrome, although all children showed difficulty compared to

						showed intersensory matching in inanimate condition.	matching the sounds and movements of objects.
Haviland et al., 1996	3- to 4-year-old typically developing (TD) children & 3- to 20-year-olds with autism	Two identical women side-by-side each speaking in a different emotion (happy, sad, angry, fearful) in German in synchrony with each other. Both events were played with a soundtrack matching one of the emotions.	2-screen intermodal preference method	Affect	PTLT to the facial information matching the vocal information	Children with autism looked significantly less to emotion events than TD peers. However, all children looked to the facial information matching the vocal information significantly more than chance for fear.	Children with autism showed some evidence of intersensory matching of faces and voices on the basis of affect.
Bebko et al., 2006	4.1- to 6.5-year-olds with autism or developmental disability & 1.7- to	Two identical women side-by-side each counting and then each speaking in synchrony with	2-screen intermodal preference method	AV temporal information	PTLT to the mouth movements matching the vocal information	Children with autism did not look to the mouth movements matching the vocal	Children with autism show a deficit in intersensory matching of faces and voices

	4.2-year-old TD children	each other. Both events played in synchrony with a soundtrack matching one of the sentences.				information significantly more than chance, whereas TD children did.	compared to TD children
Kahana-Kalman & Goldman, 2007	26- to 71-month-olds with autism & 24- to 79-month-old TD children	Two identical women (child's mother or unfamiliar woman) side-by-side each speaking in a different emotion (happy, sad, angry) in synchrony with each other. Both events were played 5-s out-of-synchrony with a soundtrack that matched one of the emotions.	2-screen intermodal preference method	Affect	Percentage of looking time to the facial movements matching the vocal information	For unfamiliar woman, children with autism did not look to the facial movements matching the vocal information significantly more than chance and looked significantly less than TD children. For mother, all children looked to the facial movements	Compared to TD children, children with autism show a deficit in matching faces and voices on the basis of affect for an unfamiliar woman, but not for their mother.

						matching the vocal information significantly more than chance	
Pons et al., 2013	4- to 7-year-olds with selective language impairment (SLI) or TD, and 3- to 6-year-olds matched for language skill	Two identical women side-by-side each speaking a different script. One event was synchronous with the soundtrack, and the other event varied by one of four levels of asynchrony: 366-ms or 666-ms where the audio preceded the video or 366-ms or 666-ms where the video preceded the audio	2-screen intermodal preference method	AV temporal information	PTLT to the mouth movements matching the vocal information	All three groups of children detected an asynchrony of 666-ms when the voice preceded lip motion. The TD and language matched children, but not children with SLI, detected the asynchrony of 666-ms when the lip motion preceded the voice. None of the groups were able to	Children with SLI showed poorer intersensory matching across face and voice on the basis of temporal information compared to TD children

						detect an asynchrony of 366-ms.	
Grossman et al., 2015	8- to 19-year-olds with high-functioning autism (HFA) or TD	Two identical women side-by-side each speaking sentences in synchrony with each other. One event was synchronous with the soundtrack and the other was asynchronous with the audio 330-ms behind the video.	2-screen intermodal preference method. Children first received implicit instructions ("look and listen") and then explicit instructions ("listen carefully and look only at the person speaking").	AV temporal information	PTLT to mouth movements matching the vocal information	All children looked to mouth movements matching the vocal information significantly more than chance after implicit and explicit instructions, but TD children looked significantly longer than children with HFA.	Children with HFA show evidence for intersensory matching of faces and voices, but spend less time looking at a synchronously speaking face than TD peers.
Patten et al., 2016	36- to 71-month-olds with ASD	Two identical women holding a doll side-by-side each telling a different story in synchrony	2-screen intermodal preference method	AV temporal information	PTLT to the mouth movements and gestures matching the	Children looked to the mouth movements and gestures matching the	Children with ASD show evidence of intersensory matching on the basis of

		with each other. Both events played in synchrony with a soundtrack matching the mouth movements and gestures of one woman while she named the doll and bounced it. The woman's face was not visible for half of the videos.			vocal information	vocal information significantly more than chance when faces were not visible, but not when they were visible.	temporal synchrony when faces are not visible, but not when faces are visible
Righi et al., 2018	5-year-olds with ASD & 3-year-old TD children	Two identical women side-by-side each speaking in synchrony with each other. One event was synchronous with the soundtrack, and the other event varied by a 0.3,	2-screen intermodal preference method	AV temporal information	PTLT to the mouth movements matching the vocal information	Children with ASD did not look to the mouth movements matching the vocal information significantly more than chance at all. TD children	Intersensory matching of faces and voices on the basis of temporal information was impaired in children with ASD.

		0.6, or 1-s asynchrony				looked to the mouth movements matching the vocal information significantly more than chance for asynchronies of 0.6 and 1-s.	
Imafuku et al., 2019	6-, 12-, and 18-month-olds born full- and pre-term	Two identical women side-by-side each reciting a different sentence from a Japanese children's story in synchrony with each other. Both events played with a soundtrack matching one of the sentences.	2-screen intermodal preference method	AV temporal information	PTLT to the mouth movements matching the vocal information	Infants born pre-term did not look at the mouth movements matching the vocal information significantly more than chance at any age. Infants born full-term looked at the mouth movements matching the vocal	Infants born pre-term did not show evidence of intersensory matching of faces and voices across the first year and a half, while infants born full-term did.

						information significantly more than chance at 6 and 18 months.	
Predicting developmental outcomes							
Righi et al., 2018	5-year-olds with ASD & 3-year-old TD children	Two identical women side-by-side each speaking in synchrony with each other. One event was synchronous with the soundtrack, and the other event varied by a 0.3, 0.6, or 1-s asynchrony	2-screen intermodal preference method	AV temporal information	Bayley Scales of Infant Development, Stanford Binet Intelligence Test, Weschler Preschool and Primary Scale of Intelligence, & Preschool Language Scales	PTLT to the mouth movements matching the vocal information in the 1-s asynchrony condition predicted receptive and expressive language scores	Greater intersensory matching across faces and voices is related to greater receptive and expressive language
Imafuku et al., 2019	6-, 12-, and 18-month-olds born full- and pre-term	Two identical women side-by-side each reciting a different sentence from a Japanese children's story	2-screen intermodal preference method	AV temporal information	Japanese MB-CDI for receptive and expressive vocabulary	PTLT to the mouth movements matching the vocal information was positively correlated with	Greater intersensory matching of faces and voices is related to greater

		in synchrony with each other. Both events played with a soundtrack matching one of the sentences.				receptive vocabulary at 12 and 18 months in infants born full- and pre-term	receptive language
Theoretical constructs: Perceptual narrowing of intersensory processing							
Lewkowicz & Ghazanfar, 2006	4-, 6-, 8-, and 10-month-olds	Two identical rhesus monkeys side-by-side each producing a sound (coo vs. grunt). Both events started in-synchrony with the soundtrack, but the offset matched only one of the sounds.	2-screen intermodal preference method	AV duration and event offset	Proportion of looking to the mouth movements matching the sound (coo vs. grunt)	Infants 4 and 6 months looked to the mouth movements matching the vocal information significantly more than chance, but infants 8 and 10 months did not.	Infants in the first half of first year showed cross-species face-voice matching on the basis of AV duration and event offset, while infants in second half of first year did not.
Pons et al., 2009	6- and 1l-month-olds. Half exposed only to Spanish and	Two identical women side-by-side, each articulating a different syllable (ba v. va) in	2-screen intermodal preference method	AV language identity	PTLT to the mouth movements matching the vocal information	At 6 months, all infants looked to mouth movements matching vocal information	By 11 months, infants exposed only to Spanish showed intersensory narrowing to matching of a

	half only English.	synchrony with each other. Both events played silently followed by an auditory-only presentation of a soundtrack of one of the syllables, and then both events played silently (removing synchrony).				significantly more than chance. At 11 months, only infants exposed to English looked significantly more than chance.	non-native language, whereas infants exposed only to English continued to match phonemes in their native language
Lewkowicz & Pons, 2013	6- to 8- and 10- to 12-month-olds exposed only to English	Two identical women side-by-side each speaking a script in a different language (English v. Spanish) in synchrony with each other. Both events played silently followed by an auditory-only	2-screen intermodal preference method	AV language identity	PTLT to the facial information matching the language	Infants 10–12 months familiarized with English looked to the speaking silently in Spanish significantly more than chance. Infants 6–8 months did not. No differences for infants	By the end of the first year, infants recognize the amodal identity of their native language, but not a non-native language

		presentation of a soundtrack of one of the languages, and then both events played silently (removing synchrony).				familiarized with Spanish.	
Kubicek et al., 2014	4.5-, 6-, and 12-month-olds exposed only to German	Two identical women side-by-side each reciting a nursery rhyme in a different language (German v. French) in synchrony with each other. 4.5- and 6-month-olds saw both events played silently followed by an auditory-only presentation of a soundtrack of one of the languages, and	2-screen intermodal preference method	1. Asynchrony: AV language identity 2. Synchrony: temporal information	PTLT to the facial information matching the language	In absence of synchrony, infants 4.5 months matched German and French, but infants 6 months only matched native German. In presence of temporal synchrony, infants 6 months matched German and French, but infants 12 months only	Infants 4.5 months can match faces and voices of native and non-native speech in the absence of temporal synchrony. Infants 6 months can only match native speech in the absence of temporal synchrony, but were able to match both native and non-native speech in presence of

		then both events played silently (removing synchrony). 6- and 12-month-olds saw both events played in synchrony with a soundtrack matching one of the languages.				matched non-native French.	synchrony. Infants 12 months only matched non-native speech in the presence of synchrony.
Lewkowicz et al., 2015	4-, 8- to 10-, and 12- to 14-month-olds exposed to English language 81% time or more	Two identical women side-by-side each speaking a different monologue in the same language (English or Spanish) in synchrony with each other. Both events played in-synchrony with a soundtrack matching one of	2-screen intermodal preference method. Half of the infants received English videos and the other half received Spanish.	1.Synchrony: AV temporal information 2. Asynchrony: AV prosody	PTLT to the facial information matching the monologue	Infants 4 and 8–10 months did not look to the facial information matching the monologue significantly more than chance for English, but infants 12–14 months did for both the native and non-native languages presented in-	By the end of the first year, infants match audiovisual native speech on the basis of prosody, but match non-native speech on the basis of temporal synchrony

		the monologues. A subgroup (12- to- 14-month-olds) also saw both events played 666-ms out-of-synchrony.				synchrony, and only for the native language presented out-of-synchrony.	
Shaw et al., 2015	5- to 10-month-olds exposed only to English	Two identical women side-by-side each speaking the same story in a different language (English v. Spanish) in synchrony with each other. Featured only bottom half of face. Both events played in synchrony with a soundtrack matching one of the languages.	2-screen intermodal preference method	AV language identity	PTLT to the facial information matching the language	Infants 5 months looked to the facial information matching language similarly for English and Spanish. Closer to 10 months, infants looked to the facial information matching the language significantly more for English than Spanish.	Infants look to the facial information matching a spoken language increasingly more for their native language and increasingly less for a non-native language across the second half of the first year

Table 2

Summary of studies and findings for intersensory processing of faces and voices as assessed by the habituation method.

Reference	Participant /Age	Stimuli: Habituation	Stimuli: Test Trials	Variables Manipulated	Research Question	DV	Results	Conclusions
Detection of global amodal properties								
Lewkowicz, 2000	4-, 6-, and 8-month-olds	AV dynamic synch: A woman repeating a syllable (ba or sha) in adult-directed manner	AV dynamic synch or asynch: Change in syllable with no change in speech OR new syllable with new speech type OR same woman speaking 666-ms out of synch OR new woman speaking	Infant age (4 vs. 6 vs. 8 months) & featural information (familiar vs. novel woman) & speech attributes (adult- vs infant-directed)	Can 4-, 6-, 8-, or 10-month-old infants discriminate speech-sound changes better in infant- or adult-directed speech? Can they discriminate familiar vs. novel people speaking out of synchrony?	Visual recovery to the switches in prosodic, temporal, and featural information	For synch, infants 4 and 6 months showed recovery for new syllable in adult-directed speech, whereas all infants did for infant-directed speech. For asynch, 4-month infants did not show recovery for familiar person, but did for novel person,	Infants can discriminate AV attribute changes in in infant-directed speech. Infants discriminate featural changes prior to temporal changes, but discriminate both by 6 months.

			666-ms out of synch				whereas 6- and 8-month infants showed recovery for both.	
Lewkowicz, 2003	4-, 6-, 8-, and 10-month-olds	AV dynamic synch: A woman repeating a syllable (ba) in a rhythmic or nonrhythmic fashion	AV dynamic synch or asynch: No change from habituation OR novel rhythm in synchrony OR novel rhythm 666-ms out of synchrony OR no rhythm 666-ms out of synchrony	Temporal information (synch vs. 666-ms asynch) & rhythm (with vs. without rhythm) & infant age (4 vs. 6 vs. 8 vs.10 months)	Can 4-, 6-, 8-, or 10-month-olds discriminate a change in temporal information (synch vs. asynch) with a change in rhythm (with vs. without) in dynamic, AV displays?	Visual recovery to the switch in rhythm and temporal information	All infants showed recovery to the synch novel rhythm, but only 10-month infants showed recovery to asynch novel rhythm.	Across the first year, infants can discriminate between AV rhythms, but can only discriminate between AV asynchronous rhythms later in the first year.

Lewkowicz, 2010	4-, 6-, 8-, and 10-month-olds	AV dynamic synch or asynch: A woman repeating a syllable (ba) in synchrony or 666-ms out of synchrony in natural speech or with a tone	AV dynamic synch or asynch: No change from habituation or change in temporal information (0-, 366-, 500-, or 666-ms out of synchrony)	Temporal information (0-, 366-, 500-, or 666-ms asynchronies) & speech type (natural vs. tone) & infant age (4 vs. 6 vs. 8 vs.10 months)	Can 4-, 6-, 8-, or 10-month-olds discriminate a change in temporal information in natural speech or in speech that removes the spectral information (tone)?	Visual recovery to the switch in temporal information	Following habituation to synch syllable presented in natural speech or as a tone, all infants showed recovery to the largest asynchrony (666-ms). Following habituation to 666-ms asynchrony, all infants showed recovery to largest (0-ms) and a smaller (366-ms) asynchrony.	Infants discriminate changes in temporal information with and without spectral information.
Pons & Lewkowicz, 2014	8-month-old infants exposed to Spanish/Catalan	AV dynamic synch: A woman speaking a script in	AV dynamic synch or asynch: No change from	Temporal information (366- vs. 500- vs. 666-ms	Can 8-month-old infants discriminate 366- vs. 500- vs.	Visual recovery to the switch in temporal information	Regardless of language, infants showed recovery to largest (666-	Infants discriminate between temporal changes in familiar and

		infant-directed speech. Half received script in Spanish and other half received English.	habituation or change in temporal information (366-, 500-, or 666-ms out of synchrony)	asynchronies)	666-ms asynchronies in dynamic, AV displays?		ms) and middle (500-ms) asynchronies.	unfamiliar languages
Detection of attributes defined by a combination of amodal properties								
Walker-Andrews & Grolnick, 1983	3- and 5-month-olds	AV static: A picture of a woman depicting a sad or happy facial expression accompanied with a soundtrack matching the facial expression	AV static: No change from habituation or change in vocal affect	Affect (happy vs. sad) & infant age (3 vs. 5 months)	Can 3- or 5-month-old infants discriminate happy vs. sad vocal affect in static, AV displays?	Visual recovery to the switch in emotional information	5-month infants showed recovery when vocal expression changed, but only 3-month infants who received a change from sad to happy showed recovery	At 5 months, infants consistently discriminate audiovisual emotional expressions, but at 3 months only discriminate a change from sad to happy
Caron et al., 1988	4-, 5-, and 7-month-olds	AV dynamic synch: A woman speaking a	AV dynamic synch: Novel woman	Affect (happy vs. sad) & infant age (4	Can 4-, 5-, or 7-month-old infants discriminate happy vs.	Visual recovery to the switch in	Infants 5 months showed recovery for both emotion	Infants can consistently discriminate audiovisual emotional

		script in a happy or sad manner for 4- and 5-month olds or a happy or angry manner for 5- and 7-month-olds	presenting same expression or novel woman presenting novel expression	vs. 5 vs. 7 months)	sad affect in dynamic, AV displays?	emotional information	changes, but infants 4 months only did for sad to happy. Infants 7, but not 5 months showed recovery to the happy to angry emotion change.	expressions as early as 5 months.
Bahrick & Pickens, 1988	5-month-olds	AV dynamic synch: A woman reciting one of two passages in English or Spanish	AV dynamic synch: No change from habituation, a change in passage with a change in language, or a change in passage with habituated language	Language (English vs. Spanish)	Can 5-month-old infants discriminate speech in English vs. Spanish in dynamic, AV displays?	Visual recovery to the switch in passage and language type	Infants showed recovery to the novel passage in a novel language, but not a novel passage in habituated language.	Infants can discriminate speech on the basis of language membership.

Walker-Andrews & Lennon, 1991	5-month-olds	AV static: A picture of a woman depicting a happy or angry facial expression accompanied with a matching or mismatching soundtrack depicting a happy, sad, or angry vocal expression	AV static: No change from habituation, or change from matching to mismatching vocal expression, mismatching to matching vocal expression, mismatching to different mismatching vocal expression	Vocal-emotional expression pairing	Can 5-month-old infants discriminate a change in vocal-emotional pairings in static, AV displays?	Visual recovery to the switch in emotional information	Infants showed recovery when they received a change from happy to sad/angry and angry to happy/sad.	Infants can discriminate audiovisual emotional expressions.
Gogate & Bahrick, 1998	7-month-olds	AV, dynamic synch: A woman speaking vowels (a, i) while	AV, dynamic synch: No change from habituation followed	Arbitrary vowel-object pairings	Can 7-month-olds detect arbitrary vowel-object relations?	Visual recovery to the switch in vowel-object pairings	Only infants who received the object moving in synchrony with the vowel showed	Infants discriminate change in arbitrary vowel-object relations, but only in the

		moving an object synchronously or asynchronously with the vowel or keeping the object still	by a switch in the vowel-object pairings		Does intersensory redundancy facilitate this discrimination?		recovery to the change in vowel-object pairings	context of intersensory redundancy
Gogate & Bahrick, 2001	7-month-olds	AV, dynamic synch: A woman speaking vowels (a, i) while moving an object synchronously or asynchronously with the vowel or keeping the object still	AV, dynamic synch: No change from habituation followed by a switch in the vowel-object pairings. Followed by memory test 10 minutes or 4 days after.	Arbitrary vowel-object relations	Can 7-month-olds detect and remember arbitrary vowel-object relations? Does intersensory redundancy facilitate this discrimination?	Visual recovery to the switch in vowel-object pairings	Only infants who received the object moving in synchrony with the vowel showed recovery to the change in vowel-object pairings. These infants showed memory for these relations 10 minutes and 4 days after habituation.	Infants discriminate and show memory for arbitrary vowel-object relations, but only in the context of intersensory redundancy

Bahrick et al., 2005	2-, 4-, and 6-month-olds	AV dynamic synch: Half saw a video of a woman or a man speaking a nursery rhyme. The other half saw a video of the woman or man speaking the nursery rhyme paired with a soundtrack of another woman or man.	AV dynamic synch: Switch in the face-voice pairings	Arbitrary face-voice relations & infant age (2 vs. 4 vs.6 months)	Can 2-, 4-, or 6-month-olds discriminate a change in face-voice pairings in dynamic, AV displays?	Visual recovery to the switch in face-voice pairings	4- and 6-month (but not 2-month) infants showed recovery to the change in the audiovisual face-voice pairings.	Infants 4 and 6 months can discriminate arbitrary intermodal relations between the appearance of a face and the particular sound of a voice
Flom & Bahrick, 2007	3-, 4-, 5-, and 7-month-olds	AV dynamic synch or asynch: A woman speaking a	AV dynamic synch or asynch: Change in affective	Affect (happy vs. angry vs. sad) & infant age (3 vs. 4 vs. 5	Can 3-, 4-, 5-, or 7-month-old infants discriminate happy vs.	Visual recovery to the switch in affect	Infants 4, 5, and 7 (but not 3) months showed recovery to a change in	By 4 months, infants can discriminate a change in audiovisual synchronous

		script in a happy, angry, or sad affect	expression for both groups	vs. 7 months)	sad vs. angry affect in dynamic, AV displays? Is this discrimination facilitated by intersensory redundancy ?		affect, but only when presented in synchrony	affect. By 5 months, synchrony is not necessary for affect discrimination
Vaillant-Molina & Bahrick, 2012	5.5-month-olds	AV dynamic synch OR V-only, dynamic: A woman eliciting and speaking in an affective expression (happy/ excited or fearful/ avoidant), each associated	AV dynamic synch OR V-only, dynamic: Switch in relation between toy and affective expression for both groups	Affect-object pairing	Can 5.5-month-old infants discriminate a change in affect-object pairings in dynamic, AV displays? Does intersensory redundancy facilitate this	Visual recovery to the switch in the affect-object pairing	Infants showed recovery to a change in the affect-object pairing for audiovisual synchrony, but not for unimodal visual.	Infants discriminate changes in affect-object pairings, but only in the context of intersensory redundancy.

		with a different moving toy.			discriminati on?			
Flom et al., 2018	3- and 5-month-olds	AV dynamic synch: A male or female 4-month old infant conveying a positive or negative affect	AV dynamic synch: Change in affective expression (positive to negative or vice versa)	Habituation time (50% standard vs. 70% extended) & infant age (3 vs. 5 months)	Can 3- vs. 5-month-old infants discriminate positive vs. negative affect of their peers in dynamic, AV displays? Do 3-month-olds require more familiarizati on than 5-month-olds?	Visual recovery to the switch in infant affect	With standard (50%) habituation criterion, 5- but not 3-month-old infants showed recovery to the change in affect. With longer (70%) criterion, 3-month-olds also showed recovery to the change in affect	Infants can discriminate the affect of their peers, and younger infants require longer habituation times.
Bahrick et al., 2019	4-month-olds	AV, dynamic synch: A woman speaking one of two passages in one of two	AV, dynamic synch: Same woman speaking novel passage in	Prosody (approval vs. prohibition)	Can 4-month-old infants discriminate approval vs. prohibition of a woman in dynamic,	Visual recovery to the switch in prosody	Only infants who received the bimodal synchronous passage showed recovery to	Infants discriminate a change prosody, but only in the context of intersensory redundancy.

		prosodic patterns (approval or prohibition). Assigned to bimodal synchronous, unimodal auditory, or bimodal asynchronous condition)	same prosody OR the same woman speaking novel passage in novel prosody		AV displays? Does intersensory redundancy facilitate this discrimination?		the change in prosody	

Table 3

Summary of studies and findings for intersensory processing of faces and voices as assessed by the McGurk task.

Reference	Participants /Ages	Stimuli	Method	DV	Results	Conclusions
Infant perception of McGurk effect						
Burnham & Dodd, 1996	17- to 20-week-olds	A soundtrack of a woman speaking a syllable (ba) paired with a live woman articulating a matching (ba) or a mismatching (ga) syllable.	Habituation: half to matching and the other half to mismatching audiovisual vowel. Test trials were [ba] or fused [da]	Fixation duration during test phase	Infants habituated to the mismatching audiovisual vowel spent a significantly longer duration fixating the fused percept in the test trials.	Infants 4.5–5 months perceive the McGurk effect
Rosenblum et al., 1997	5-month-olds	A male uttering audiovisual congruent [va], incongruent [ba-va], or incongruent da-va] in videos	Habituation: habituated to audiovisual congruent [va]. Test stimuli were incongruent audiovisual [ba-va] and [da-va]	Looking time to incongruent test trials	Looking time to incongruent test trials significantly differed from looking to final congruent habituation trial for incongruent [da-va], but not incongruent [ba-va] trials	Infants 5 months demonstrate McGurk-life effect, indicated by their looking to incongruent audiovisual [da-va] trials (portraying the McGurk/fusion effect)

Burnham & Dodd, 2004	17- to 19-week-olds	Audio track of women saying [ba], [ga], and [da]. Mimers mimed the syllables in synchrony with audio track in real-time	Habituation: Half habituated to incongruent (McGurk) stimuli and half habituated to congruent audiovisual [ba]. Test trials were of each vowel.	Visual fixation time to mimer face for habituation and test trials	Infants habituated to incongruent stimuli showed greater recognition for fused/McGurk vowel [da] than infants habituated to audiovisual congruent [ba] vowel.	Infant behavior is consistent with perception of the McGurk effect, similarly to adults.
Desjardins & Werker, 2004	4-month-olds	A woman speaking [vi], [bi], and [shu] in videos	Habituation: infants were habituated to either audiovisual congruent [vi] or [bi] and dishabituated to incongruent [bi-vi] or [vi-bi].	Looking time during dishabituation trials	For infants habituated to audiovisual congruent [bi], females looked longer to incongruent dishabituation [bi-vi], but not [vi-bi] trials. For infants habituated to [vi], neither sex looked longer to incongruent dishabituation [bi-vi], but not [vi-bi] trials, suggesting they perceived	Infants integrate audiovisual speech and perceive McGurk effect under some conditions, with female infants perceiving it more.

					[bi-vi] as fused [vi].	
Kushnerenko et al., 2008	20- to 23-week-olds	A woman saying [ba] and [ga] in congruent audiovisual speech and incongruent [ga-ba and ba-ga] audiovisual speech in videos	EEG was used while infants watched videos for 4–9 minutes	Occipital, temporal, and frontals ERPs	[ba-ga] was processed as mismatched, indicated by an additional activation starting at 290-ms from sound onset over frontal and temporal areas. The incongruent McGurk stimulus pair [ga-ba], was not processed as a conflicting stimulus by the infants, as indicated by the lack of ERP difference when compared to the congruent [ba] and [ga] syllables	Infants failed to detect the mismatch in the McGurk stimulus, suggesting they perceived it as the fused percept [da].
Child perception of the McGurk effect						
McGurk & MacDonald, 1976	3- to 4-year-olds, 7- to 8-year-olds,	A woman saying [ba],	Auditory-only tracks of syllables were	Verbal response of perceived	Auditory accuracy was higher than	First study to demonstrate the role of vision in

	and 18- to 40-year-olds.	[ga], [pa], or [ka] in videos	played followed by videos with four audiovisual combinations: [ba-ga], [ga-ba], [pa-ka], [ka-pa]	syllable spoken by woman	audiovisual accuracy for every age. Audiovisual McGurk illusions were observable across all ages.	the perception of speech via the McGurk effect.
Massaro, 1984	5- to 11-year-olds and adults	A video of a man mouthing the syllables [ba] and [da] or keeping his mouth closed played with a soundtrack of synthetic speech producing a range of five syllables gradually changing from [ba] to [da]	All trials were audiovisual: 3 levels of visual information combined with the 5 levels of auditory information	Forced 2-choice button press for perceived syllable	Children were influenced by the visual information significantly less than adults.	Child perception of audiovisual speech is less influenced by visual information relative to adults.
Dupont et al., 2005	4- to 5-year-olds and 22- to 31-year-old women	The lower half of a speaker's face saying [aba], [ada], [aga], [ava], [ibi], [idi],	Videos were played in 4 conditions: bimodal non-conflicting, visual-only,	Verbal response of perceived sequence	Child performance in visual-only was significantly worse than adult performance and	Children showed the McGurk effect some of the time, but were more sensitive to auditory

		[igi], and [ivi] three times each on a video	auditory-only, and bimodal conflicting		child performance for auditory only and non-conflicting sequences. For conflicting sequences, children reported sequences as the auditory sequence (opposed to visual) significantly more than adults. Child perception of McGurk illusion was not significantly different from adults.	information, reporting a higher number of audio sequences in response to the conflicting sequences
Tremblay et al., 2007	5- to 9-, 10- to 14-, and 15- to 19-year-olds	A male speaking syllables [ba], [ga], and [va] in videos	Syllables were played in auditory-only, audiovisual congruent, audiovisual incongruent [ba-ga], and visual-only	Verbal response of syllable heard	For auditory-only, audiovisual congruent, and visual-only conditions, performance was similar across ages. For audiovisual	Children 5–9 years do not perceive the McGurk illusion at the same rate as older children and adolescents

					incongruent condition, children 5 to 9 years perceived significantly less McGurk illusions than older ages	
Sekiyama & Burnham, 2008	6-, 8-, 11-, and 18- to 29-year-olds. Half spoke English and half spoke Japanese	Two males and two females (one each English and Japanese) speaking [ba], [da], and [ga] in videos.	Syllables were played in auditory-only, visual-only, audiovisual congruent, and audiovisual congruent conditions at 4 levels of signal-to-noise ratios (SNRs): no noise, -4, +4, and +12 decibels. Half received increase in SNR and half received decrease.	Forced 3-choice button press for perceived syllable	Children 6 years showed weak perception of McGurk illusion, regardless of language spoken. By 8 years, English speaking children showed increase in perception of McGurk illusion, but Japanese speaking children did not. Perception of McGurk illusion was larger at lower SNRs.	Inter-language differences in perception of McGurk effect emerge between 6 and 8 years.
Nath et al., 2011	6- to 12-year-olds	A woman speaking [ba],	Syllables were played	Behavioral was verbal	59% of children perceived the	There are individual

		[ga], [da], and [ma] in videos	auditory-only, audiovisual congruent, McGurk incongruent [ba-ga], and non-McGurk incongruent [ga-ba]. Assessed behaviorally and FMRI. FMRI added in audiovisual [ma]	response of syllable heard. FMRI was button press for each audiovisual [ma]	McGurk effect and 41% did not. The left superior temporal sulcus (STS) and the left and right fusiform gyri were more active in children that perceived the McGurk effect.	differences among children 6–12 years in perception of the McGurk effect. Increased activity in the left STS was observed for children that perceived the McGurk effect.
Hirst et al., 2018	3- to 6-, 7- to 9-, 10- to 12-, and 20- to 35-year-olds	A woman speaking [ba], [ga], and [da] in videos	Syllables were played audiovisual congruent and audiovisual incongruent [ga-ba]. Played with 5 levels of visual noise and also with 5 levels of auditory SNRs: no noise, -2, -8,	Forced 3-choice key press of heard syllable	Children 3 to 6 and 7 to 9 years made significantly less McGurk responses than adults. For auditory noise, 3- to 6- and 7- to 9-year-olds both required significantly more noise to induce McGurk effect	The influence of vision over audition, and thus the susceptibility to the McGurk effect, increases across development

			-14, and -20 decibels		than adults. For visual noise, adults required significantly more noise to eliminate McGurk responses compared with 3- to 6-year-olds.	
Perception of the McGurk effect in infants at risk for autism and children displaying developmental disabilities						
Hayes et al., 2003	8- to 14-year-old TD and learning-disabled children	A woman speaking [ata], [apa], and [aka] in videos.	Syllables were played auditory-only, visual-only, and audiovisual incongruent at 3 SNRs: quiet, 0, and 212 decibels	Verbal response of perceived word	Both groups of children reported more visual responses as SNR increased. At the highest SNR, learning disabled children reported significantly more visual responses, while TD children reported significantly more McGurk responses	When presented with incongruent audiovisual speech, learning disabled children were more likely to report the visual component, while TD children were more likely to report combination and McGurk responses.
Williams et al., 2004	5- to 13-year-old TD children or	A computer generated face articulating the	Syllables were played auditory-only, visual-only,	Verbal response of perceived syllable	After controlling for performance in unimodal trials, children with	Children with ASD show similar intersensory processing of

	children with ASD	syllables [ba], [tha], and [da].	audiovisual congruent, and audiovisual incongruent		ASD did not significantly differ from TD children in accuracy of syllable identification.	audiovisual speech events when compared to TD children, after accounting for unimodal performance.
Norrix et al., 2007	49- to 70-month-olds with SLI and 51- to -70-month-olds with normal language skills	Children speaking [bi], [gi], [di] and [thee] in videos. [bi] and [gi] were paired with pictures of a bee and gee (karate uniform). [di] and [thee] were McGurk stimulus paired with a picture of twins Dee and Thee	Syllables were played auditory-only, audiovisual congruent, and audiovisual incongruent	Verbal 3-choice forced response from the 3 pictures. Expressive vocabulary from the Structured Photographic Expressive Language Test	Children with SLI reported the McGurk effect significantly less than children with normal language skills. Greater integrated responses were associated with higher expressive vocabulary size	Children with SLI are less influenced by visual information from audiovisual speech than children with normal language skills
Dodd et al., 2008	Experiment 1: 38- to 67-month-old children with speech delay	Experiment 1: a man and woman saying [ba], [ga], and [da] in 2 side-	Experiment 1: syllables were presented visual-only, auditory-only,	Experiment 1: verbal forced choice Y/N response if the man and lady	Experiment 1: children with speech delay and TD children did not differ	Children with speech delay do not perceive McGurk effect different from TD

	or TD. Experiment 2: 3- to 5-year-olds with phonological delay or disorder	by-side videos. Experiment 2: a woman with Dublin accent speaking names of 6 pictures: pea, tea, key, bow, dough, go.	audiovisual congruent, audiovisual McGurk [ba-ga], and audiovisual combination [ga-ba]. Experiment 2: words were presented in audiovisual congruent and incongruent	said the same word. Experiment 2: forced 6-choice point response to picture	significantly in perception of the McGurk illusion. Experiment 2: children with phonological delay were more likely to report the visual component of the illusion compared to children with phonological disorder.	children, but children with phonological delay perceive McGurk effect different from children with phonological disorder
Mongillo et al., 2008	8- to 19-year-olds with ASD and 11- to 19-year-old mental age matched TD children	A woman speaking syllables [ba], [da], [va], and [tha]	Syllables were presented in audiovisual congruent and audiovisual incongruent (auditory ba combined with visual da, va, and tha)	Forced 2-choice button press of perceived syllable	Children with ASD perceived the McGurk effect significantly less than TD children.	Children with ASD are less influenced by the visual information in audiovisual speech than TD children.
Taylor et al., 2010	7- to 16-year-olds with autism & 8- to 16-year-	A woman speaking [aba], [ava], [atha], [ada], [aga] in videos	Words were presented in auditory-only, visual-only, audiovisual	Auditory accuracy, visual accuracy, and McGurk effect	Children with autism perceived McGurk effect significantly less than TD children	At 7 years, children with autism perceived the McGurk effect less than 8-year-

	old TD children		congruent, and audiovisual incongruent		at youngest ages, but children with autism showed a faster rate of development in perception of McGurk effect relative to TD children, with no significant differences between groups at oldest ages tested	old TD children, but by 16 years, children with autism perceived the McGurk effect similarly to TD children
Iarocci et al., 2010	10.6-year-olds with autism & 10.3-year-old TD children	A male's mouth and nose region speaking [ba], [tha], [va], and [da] in videos	Syllables were presented auditory-only, visual-only, audiovisual congruent, and audiovisual incongruent	Verbal response of perceived syllable	Children with autism made significantly less visually compatible responses than TD children. For both groups, poorer visual-only performance was associated with lower visual influence on speech perception (and vice versa).	Poor lip-reading (visual-only performance) influences audiovisual speech perception in children with and without autism.

Irwin et al., 2011	5- to 15-year-olds with autism & 7- to 12-year-old TD children	A male speaking [ma], [na], [ga], and [ba-da] in videos. There were 6 SNRs: 5, 0, -5, -10, -15, -20 decibels. Non-speech stimuli were eight shapes changing size with sine wave tones.	Syllables were presented visual-only, auditory-only, audiovisual congruent, audiovisual incongruent [ga-ma], and 4 audiovisual asynchronous conditions (visual or auditory lead at 250 or 550-ms)	Verbal response of heard syllable. For audiovisual asynchronous, judged whether it was synchronous or not.	Children with autism reported the McGurk effect significantly less than TD children. For audiovisual asynchrony, both groups performed significantly better with large (550-ms) compared to small (250-ms) asynchronies. Children with autism did not differ significantly from TD children for non-speech stimuli.	Children with autism show impairment in audiovisual speech perception, but not non-speech perception
Guiraud et al., 2012	9-month-olds at high- and low-risk for autism	Women speaking [ba] and [ga] in two side-by-side videos. The congruent face was paired	2-screen intermodal preference method	Total fixation time to eyes, mouth, and face	Infants at low-risk for autism looked significnalty longer at mouth of incongruent face in the mismatch	Infants at low-risk for autism can detect audiovisual incongruencies (demonstrated by looking at mouth longer for

		with mismatch [ba-ga] or fusion [ga-ba]			condition, but equally long at mouths of congruent and incongruent faces in the fusion condition. Infants at high-risk for autism did not show differential looking at mouth for any condition compared to the congruent face.	mismatch), but also perceive the McGurk effect (demonstrated by equal looking at mouths for fusion condition). Infants at high-risk for autism show deficits in audiovisual matching.
Woynaroski et al., 2013	8- to 17-year-olds with autism or TD	A woman speaking [ba], [ga], [da], and [tha] in videos	Syllables were presented visual-only, auditory-only, matched audiovisual, and mismatched audiovisual [ba-ga]. For mismatch, visual preceded auditory by 0, 33, 66, 100, 166, 233, and 300-ms	Percent accuracy in each condition & proportion of trials the McGurk effect was perceived; ADOS & Sensory Profile Caregiver Questionnaire	Children with autism were significnalty less accurate at identifying matched audiovisual speech and had a marginally larger temporal binding window for mismatched audiovisual speech. Children with autism who	Children and adolescents with autism show reduced multisensory speech perception compared to TD children, and this is related to characteristics of autism.

					perceived the McGurk effect less showed greater sensitivity to sound and greater difficulty functioning in presence of distractions	
Stevenson et al., 2014	6- to 18-year-olds with autism or TD	A woman speaking [ba] and [ga] in videos	Syllables were presented visual-only, auditory-only, congruent audiovisual, and McGurk audiovisual	Forced 7-choice keyboard press response of what syllable the speaker said	Adolescents with autism 13 to 18 years reported the McGurk effect significantly less than TD children. Children 6 to 12 years with autism did not differ significantly from TD children 6 to 12 years or from adolescents with autism 13 to 18 years	Children and adolescents with autism fail to show the developmental growth in audiovisual integration that is seen in TD children and adolescents
Feldman et al., 2019	8- to 17-year-olds with autism or TD	A woman speaking [ba], [ga], [da], and [tha] in videos	Syllables were presented visual-only, auditory-only, congruent	Percent correct identification for each condition, except	Children with autism showed reduced accuracy for congruent audiovisual	Reduced accuracy and poorer temporal acuity in audiovisual speech identification are

			audiovisual, and incongruent audiovisual. [ba-ga] was manipulated so visual preceded auditory by 33, 66, 100, 166, 233, and 300-ms.	incongruent audiovisual, where rates of reported McGurk effect were calculated at each asynchrony (33, 66, 100, 166, 233, and 300-ms) to derive TBW; Sensory Experiences Questionnaire & Sensory Profile Caregiver Questionnaire	speech and wider TBWs for mismatched audiovisual speech compared to TD children. Wider TBW was associated with more atypical patterns of sensory responsiveness.	related to greater atypical sensory responsiveness in children with autism 6–18 years
Perception of McGurk effect in relation to developmental outcomes						
Boliek et al., 2010	6- to 9- and 10- to 12-year-olds with learning disability or TD	A woman speaking [bi] and [gi] in videos. Soundtrack of male speaking [bi] and [gi]	Syllables were presented congruently (female face and female voice) and incongruently (female face	Forced 6-choice verbal response of perceived syllable from paper; achievement scores (math	Children with learning disability who reported the McGurk effect less had lower reading or math achievement scores (6 to 9	Children with learning disability show a weaker McGurk effect than TD children, and this is related to achievement and IQ

			and male voice) with matching and mismatching [bi-gi] phonetic information	and reading) and full-scale IQ	years) and lower IQ scores (10 to 12 years). TD children who reported the McGurk effect more had greater achievement scores.	
Barutchu et al., 2019	7- to 13-year-olds	A woman speaking [ba], [da], and [ga] in videos.	Syllables were presented audiovisual matching or mismatching [ga-ba]	McGurk response (FIND) & full-scale IQ and the Test of Everyday Attention for Children (TEA-Ch) for visual, auditory, and divided audiovisual attention	Perception of the McGurk effect was significantly related to audiovisual attention on the TEA-Ch test	Perception of the McGurk effect is related to dual attention abilities in children

Table 4

Summary of studies and findings for intersensory processing of faces and voices as assessed by the speech-in-noise task.

Reference	Participants/ Ages	Stimuli	Method	DV	Results	Conclusions
Erber, 1971	9- to 12-year-old normal hearing, severely impaired hearing, and deaf children	A woman speaking 240 common nouns at 8 SNRs in videos	Words were presented visual-only, auditory-only, and audiovisual at each SNR	Written response of word presented	Each group's performance improved when presented with words audiovisually. Hearing-impaired children required a greater SNR for word intelligibility than normal hearing children required	Word intelligibility improves when presented with words audiovisually compared to visual- or auditory-only, but the SNR required for intelligibility differs based on hearing status
Ross et al., 2011	5- to 14- and 16- to 56-year-olds	A woman speaking 300 mono-syllabic words in videos at 6 level of pink noise (no noise, 53, 56, 59, 62, and 65 decibels)	Words were presented visual-only, auditory-only, and audiovisual	Verbal report of word heard; audiovisual gain (audiovisual minus auditory-alone difference score)	Audiovisual performance increased across age. Audiovisual gain significantly increased from 5-7 to 8-9 years, but not from 8-9 to 10-11 years, and by 12-14 years had approached adult levels	Audiovisual gain in speech intelligibility increases across 5– 56 years
Knowland et al., 2016	4- to 11-year-old children with	A woman or male speaking	Sentences were presented	Forced 4-choice picture	Audiovisual performance did not significantly differ	Children with LLI do not show deficits in identifying words in

	learning language impairment (LLI) and 5- to 11-year-old TD children	sentences in English presented in noise at a level to elicit 70.7% accuracy for each individual child	auditory-only and audiovisual	response of the target of the sentence	between children with LLI or TD children. Accuracy of sentence target identification increased across age for both groups.	sentences presented in noise compared to typically developing peers
Stevenson et al., 2017	6- to 18-year-olds with autism or TD	A woman speaking 216 tri-phonemic nouns at 4 SNRs (0, -6, -12, and -18 decibels) in videos	Words were presented visual-only, auditory-only, and audiovisual at each SNR	Whole word and phoneme recognition accuracy by typing word heard on a keyboard	For both phoneme and whole-word recognition TD children were significantly more accurate than children with autism, and greater SNRs were related to greater accuracy	Children with autism 6–18 years show significant deficits in multisensory speech perception in noise compared to TD children

Table 5

Summary of studies and findings for intersensory processing of faces and voices as assessed by the eye-tracking method.

Reference	Participants/ Ages	Stimuli	Method	DV	Results	Conclusions
Selective attention to specific areas of speaking faces and change across development						
Lewkowicz & Hansen-Tift, 2012	4-, 6-, 8-, 10-, and 12-month-olds and adults exposed to monolingual English	A woman speaking a monologue in either English or Spanish	Eye-tracking: AOIs were eyes and mouth	PTLT to each AOI (mouth, eyes)	Infants presented with English monologue looked longer to eyes at 4 months, equally at eyes and mouth at 6 months, to mouth at 8–10 months, and equally to eyes and mouth at 12 months. Infants presented with Spanish monologue showed same pattern, except longer looking to mouth at 12 months.	At 8 months, there is a shift in attentional focus to the mouth of a speaker (native or non-native). At 12 months, this focus begins shifting back to the eyes for native speech, but stays at the mouth for non-native speech
Tenenbaum et al., 2013	6-, 9-, and 12-month-olds exposed to	A woman seated at a table speaking about	Eye-tracking: regions of interest	Gaze fixations	At all ages, infants looked more to face of woman	Infants 6, 9, and 12 months look more at a

	monolingual English	objects in front of her in videos	(ROIs) were eyes, mouth, object		than object. Infants 6 months showed no significant difference in looking to eyes or mouth, infants 9 months looked more at mouth than eyes, and infants 12 months looked more at mouth than eyes	speaking face than an object, but infants 9 and 12 months look at the mouth of a speaking woman more than the eyes
Tomalski et al., 2013	6- to 7- and 8- to 9-month-olds regularly exposed to English	A woman speaking [ba] and [ga] in videos in audiovisual incongruent [ba-ga, ga-ba] speech	Eye-tracking: AOIs were mouth, eyes, and entire face oval	Looking time to each AOI	Infants 6–7 months looked longer to mouth in the fusible incongruent (McGurk) presentation than in the non-fusible incongruent presentation, but infants 8–9 months looked at the mouth equally in both presentations	Infants 6–9 months increase attention to the mouth, but only when the auditory and visual information are in apparent conflict

Wilcox et al., 2013	3- to 4- and 9-month-olds	A woman speaking “Hey baby” and waving in a video and just clapping in another video	Eye-tracking: AOIs were eyes, mouth, and hands	PTLT to each AOI	Infants 3–4 months had similar PTLT to eyes and mouth, but greater PTLT to mouth than hands. Infants 9 months had greater PTLT to the eyes than mouth.	When presented with an audiovisual dynamic video with sparse linguistic content, infants 9 months look more to the eyes than the mouth
Pons et al., 2015	4-, 8-, and 12-month-olds exposed to monolingual Catalan/Spanish	Two women speaking a monologue in videos, one in Catalan/Spanish and one in English	Eye-tracking: AOIs were eyes and mouth	PTLT to each AOI	For native monologue (Catalan/Spanish) infants looked longer to eyes at 4 months, longer to mouth at 8 months, and equally to eyes and mouth at 12 months. For non-native (English) monologue, infants showed same pattern, except longer looking to mouth at 12 months.	The developmental pattern of shifting attention to the mouth generalizes to infants exposed to Spanish/Catalan, and replicates previous study by Lewkowicz & Hansen-Tift (2012)

Hillairet de Boisferon et al., 2017	4-, 6-, 8-, 10-, and 12-month-olds exposed to English greater than 80% of the time	A woman speaking a monologue in either English or Spanish in infant-directed or adult-directed speech. Soundtrack preceded video by 666-ms	Eye-tracking: AOIs were eyes and mouth	PTLT to each AOI	For English and Spanish speech, infants did not show differential looking to eyes and mouth at 4 and 6 months, longer looking to mouth at 8 months, and no differential looking at 10 and 12 months.	Asynchronous speech disrupts the developmental pattern of attention to the mouth and eyes. This disruption is evident at 10 months, when attention to the mouth was no longer present, unlike Lewkowicz & Hansen-Tift (2012)
Hillairet de Boisferon et al., 2018	14- and 18-month-olds	A woman speaking a monologue in either English or Spanish in either infant- or adult-directed speech. Infants saw one language in type of speech.	Eye-tracking: AOIs were eyes and mouth	PTLT to each AOI	Infants attended more to the speaker's mouth than to the eyes, regardless of speech condition (English, Spanish) at both ages, for infant-directed speech at 14 months, and for both adult-, and	Infants selectively attend more to a speaker's mouth than eyes across the second year. Extends findings from that of Lewkowicz & Hansen-Tift (2012)

					infant-directed speech at 18 months.	
Tsang et al., 2018	6- to 12-month-old monolingual English or bilingual (English/other language) exposed infants	A woman speaking in infant-directed speech	Eye-tracking: AOIs were eyes and mouth	PTLT to each AOI	Attention to the mouth relative to the eyes of a speaking face increased across 6 to 12 months.	Attention to the mouth increases across 6 to 12 months, replicating Pons et al. (2015), but not Lewkowicz & Hansen-Tift (2012)
Pons et al., 2019	12-month-old Catalan/Spanish exposed infants	A woman speaking a monologue in infant-directed speech in either Catalan/Spanish or English. Infants saw both languages.	Eye-tracking: AOIs were eyes and mouth	PTLT to each AOI	For native language, infants looked equally to eyes and mouth, but for non-native English, infants looked more to the mouth.	Infant selective attention to a speaker's face differs as a function of language familiarity, replicating Lewkowicz & Hansen-Tift (2012) and Pons et al. (2015)
Morin-Lessard, 2019	5-, 9-, 12-, 14-, 18- to 24-, 36-, and 48- to 60-month-old monolingual or	A woman reciting a passage in English, French, or Russian in	Eye-tracking: AOIs were eyes and mouth	PTLT to each AOI	For the dominant and nondominant languages, all infants 5 months looked to mouth	Selective attention to the mouth of a speaker's face increases across

	bilingual English/French children	infant-directed speech			and eyes equally, and children 9 to 60 months looked significantly longer to the mouth than eyes. When viewing an unfamiliar language (Russian), children showed equal attention to eyes and mouth across all ages.	the first 5 years for a native (dominant) and non-native (nondominant) language.
Imafuku et al., 2019	6-, 12-, and 18-month-olds	Two identical women side-by-side each reciting a different sentence from a Japanese children's story in synchrony with each other. Both events played with a soundtrack matching one of the sentences.	Eye-tracking: AOIs were eyes (congruent, incongruent) and mouth (congruent, incongruent)	PTLT to each AOI	18-month infants looked significantly more to the mouth (congruent and incongruent) than 12- and 6-month infants, and 12-month infants looked to the mouth significantly more than 6-month infants	Selective attention to the mouth of a speaking face increases across 6 to 18 months

Selective attention to speaking faces in monolingual versus bilingual children						
Pons et al., 2015	4-, 8-, and 12-month-olds exposed to monolingual Catalan/Spanish or bilingual	Two women speaking a monologue in videos, one in Catalan/Spanish and one in English	Eye-tracking: AOIs were eyes and mouth	PTLT to each AOI	Bilingual infants looked longer to the mouth at 4 months than monolingual infants. At 8 months, both bilingual and monolingual infants looked longer at the mouth than eyes. At 12 months, bilingual infants looked longer at the mouth than monolingual infants.	Bilingual infants display an earlier attention shift to the mouth than monolingual infants, paying more attention to the mouth at 4 months. This trend continues at 12 months, where bilingual infants pay more attention to the mouth than monolingual infants.
Fort et al., 2017	Monolingual Spanish/Catalan exposed infants at 15 months and bilingual Spanish/Catalan exposed infants at 15 months	A woman speaking a sentence in Spanish or Catalan	Eye-tracking: AOIs were eyes and mouth	PTLT to each AOI	At 15 months, bilingual infants looked significantly longer to mouth than the eyes during the speech event, whereas monolingual 15-	Bilingual infants show greater selective attention to the mouth of a speaking face than monolingual infants in second year of life.

					month infants marginally did.	
Tsang et al., 2018	6- to 12-month-old monolingual English or bilingual (English/other language) exposed infants	A woman speaking in infant-directed speech	Eye-tracking: AOIs were eyes and mouth	PTLT to each AOI	Attention to the mouth of a speaking face did not significantly differ between monolingual and bilingual exposed infants	Monolingual and bilingual infants show similar pattern of looking to mouth of a speaking face across 6 to 12 months.
Morin-Lessard et al., 2019	5-, 9-, 12-, 14-, 18- to 24-, 36-, and 48- to 60-month-old monolingual or bilingual English/French children	A woman reciting a passage in English, French, or Russian in infant-directed speech	Eye-tracking: AOIs were eyes and mouth	PTLT to each AOI	Attention to the mouth of a speaking face did not differ significantly between monolingual and bilingual exposed children for dominant and nondominant languages.	Monolingual and bilingual children show a similar pattern of looking to the mouth of a speaking face across 5 to 60 months
Visual attention to speaking faces in infants and children at risk for or displaying developmental disabilities						
Shic et al., 2014	6-month-old infants with symptoms of ASD, at high-risk for developing ASD	A woman reciting a nursery rhyme	Eye-tracking: AOIs were eyes, nose, mouth, rest of face, and background	PTLT to inner face (eyes, nose, mouth), PTLT to outer face (skin, hair, body), and eye-	Infants later diagnosed with ASD looked less at inner face features and more at outer face	Infants that are later diagnosed with ASD show reduced selective attention to a speaking face.

	(both atypical and TD) or at low-risk for developing ASD (both atypical and TD)			to-mouth looking ratio	features than the other groups of infants. All infants looked significantly less at the eyes relative to mouth.	When they attend to speaking faces, they attend to the outer features that are not socially informative.
Pons et al., 2018	5- to 9-year-old monolingual Spanish TD children or children with SLI (two subtypes: lexical and phonological-syntactic)	A woman reciting a monologue in Spanish	Eye-tracking: AOIs were eyes and mouth	PTLT to each AOI	TD children looked significantly longer at the mouth than the eyes, but children with SLI looked equally at the eyes and mouth. Children with lexical deficits looked more to the mouth than eyes, but children with phonological-syntactic deficits looked more to the eyes than mouth.	TD children showed greater selective attention to the mouth of a speaking face than the eyes. Children with SLI attended to different parts of a speaking face as a function of the specific subtype
Imafuku et al., 2019	6-, 12-, and 18-month-olds born full-term or pre-term	Two identical women side-by-side each reciting a	Eye-tracking: AOIs were eyes (congruent,	PTLT to each AOI	At 18 months, infants born full-term looked to the mouth of a	By 18 months, there are differences in selective

		different sentence from a Japanese children's story in synchrony with each other. Both events played with a soundtrack matching one of the sentences.	incongruent) and mouth (congruent, incongruent)		speaking face (congruent or incongruent) significantly longer than infants born pre-term	attention to the mouth of a speaking face between infants born full- and pre-term
Berdasco-Muñoz et al., 2019	8-month old infants born pre-term and 6- and 8-month-old infants born full-term matched for postnatal (8 months) and maturational (6 months) ages	A woman reciting a children's story in French and in English in infant-directed speech	Eye-tracking: AOIs were eyes and mouth	PTLT to each AOI	Full-term 6- and 8-month infants looked longer to the mouth for the non-native language and longer to the eyes for the native language. Pre-term infants did not show a difference for the native or non-native language.	Infants born pre-term differ from postnatal and maturational age peers in attending to the eyes and mouth of a speaking face.
Selective attention to speaking faces and relations to developmental outcomes						
Young et al., 2009	6-month-old infants at low or high risk for	Infant's mother speaking spontaneously,	Eye-tracking	Eye-mouth index score; Autism	There was no significant relation between face	Selective attention to the mouth of a

	autism seen longitudinally at 12, 18, and 24 months	depicting the still-face episode, and then re-engaging spontaneously on video		symptoms, motor skills, language skills, behavior	scanning and autism symptoms, motor skills, or behavior. Greater looking to the mother's mouth when she spoke was related to greater expressive vocabulary.	speaker's face is related to expressive vocabulary, but selective attention to the eyes or mouth are not related to autism symptoms.
Kushnerenko et al., 2013	6- to 9-month-old infants seen longitudinally at 14 to 16 months	A woman articulating two syllables (ba and ga) incongruently (visual ga with auditory ba and vice versa)	Eye-tracking: AOIs were eyes and mouth	PTLT to each AOI at 6 to 9 months; Auditory comprehension and expressive communication scores at 14 to 16 months	Infants who looked more to the eyes during the incongruent speech had greater auditory comprehension scores, whereas infants who looked longer to the mouth during incongruent speech had lower auditory comprehension scores	Selective attention to the eyes of a speaker's face when the auditory and visual speech do not match is related to later language skills
Tenenbaum et al., 2015	12-month-old infants seen longitudinally at	A woman speaking about one of two	Eye-tracking: AOIs were eyes, nose,	Mouth-to-eyes index at 12 months;	Infants who looked longer to the mouth of the	Selective attention to the mouth of a

	18 and 24 months	objects in front of her	mouth, target object, distractor object	Receptive and expressive vocabulary at 18 and 24 months	speaker's face at 12 months had greater expressive (but not receptive) vocabulary size at 18 and 24 months.	speaking face is related to expressive vocabulary in the second year.
Imafuku & Myowa, 2016	6- and 12-month-olds exposed to Japanese	A woman speaking a story in Japanese in synchrony and out of synchrony	Eye-tracking: AOIs were eyes, mouth, and face	PTLT to each AOI; Receptive and expressive vocabulary	6-month infants looked more to the mouth when the speech was synchronous than when it was asynchronous, but there was no significant difference for 12-month infants. Greater looking to the mouth (in and out of synchrony) at 6 months was related to greater receptive vocabulary size at 12 months.	Selective attention to the mouth of a speaking face differs as a function of age when the speech is presented in and out of synchrony. Selective attention to a speaker's mouth speaking in and out of synchrony is related to receptive vocabulary.
Hillairet de Boisferon et al., 2018	18-month-olds	A woman speaking a monologue in either English or	Eye-tracking: AOIs were eyes and mouth	PTLT to each AOI; expressive vocabulary size	Attention to a speaker's mouth was not related to	Selective attention a speaker's mouth is not related to

		Spanish in either infant- or adult-directed speech. Infants saw one language in type of speech.			vocabulary size at 18 months.	expressive vocabulary in the second year.
Tsang et al., 2018	6- to 12-month-old monolingual English or bilingual (English/other language) exposed infants	A woman speaking in infant-directed speech	Eye-tracking: AOIs were eyes and mouth	PTLT to each AOI; receptive and expressive vocabulary	Greater attention to the mouth from 6 to 12 months was related to greater expressive, but not receptive, language skills in monolingual and bilingual exposed infants	Selective attention to a speaker's mouth is related to expressive language skills across the second half of the first year.
Pons et al., 2019	12-month-old Catalan/Spanish exposed infants	A woman speaking a monologue in infant-directed speech in either Catalan/Spanish or English. Infants saw both languages.	Eye-tracking: AOIs were eyes and mouth	PTLT to each AOI; social/ communication skills on Bayley Scales of Infant Development	Greater attention to the eyes of a person speaking a native, but not a non-native language is related to greater social skills and marginally related to greater communication skills at 12 months.	Selective attention to the eyes of a person speaking a familiar/native language is related to social and communication skills at the end of the first year.

Morin-Lessard et al., 2019	5-, 9-, 12-, 14-, 18- to 24-, 36-, and 48- to 60-month-old monolingual or bilingual English/French children	A woman reciting a passage in English, French, or Russian in infant-directed speech	Eye-tracking: AOIs were eyes and mouth	PTLT to each AOI; receptive vocabulary at 9, 12, and 14 months; expressive (conceptual) vocabulary at 9, 12, 14, and 18 to 24 months	For monolingual, but not bilingual exposed infants, greater expressive vocabulary was related to greater looking to the mouth of a speaking face. For bilingual, but not monolingual exposed infants, greater receptive vocabulary was marginally related to less looking to the mouth of a speaking face.	Selective attention to the mouth of a speaking face is positively related to expressive vocabulary for monolingual exposed infants, and negatively related to receptive vocabulary for bilingual exposed infants.

Figures

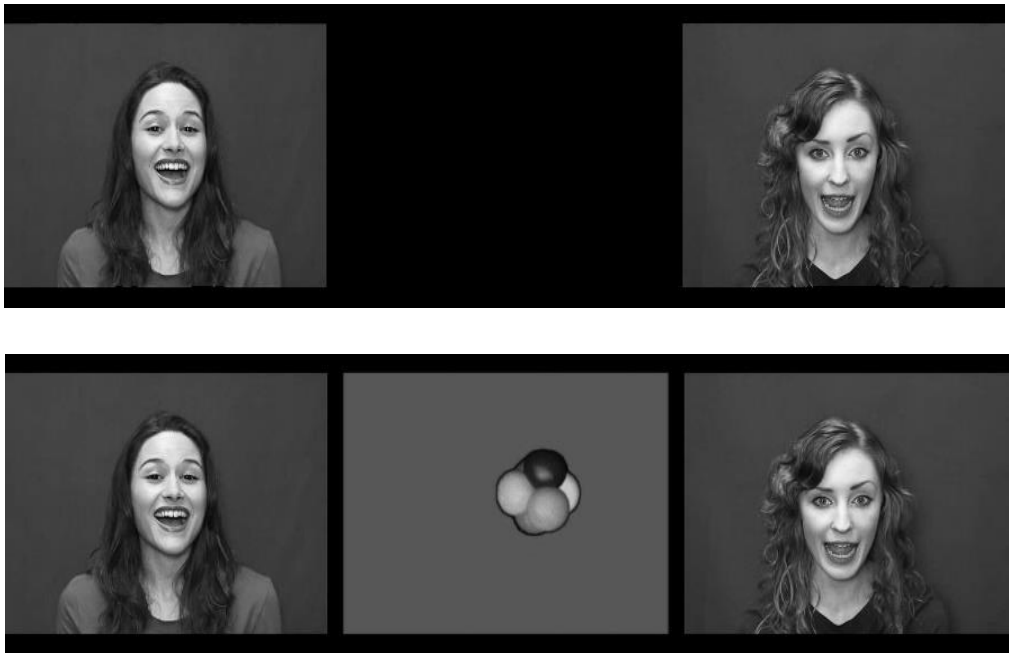

Figure 1. Static images of the dynamic audiovisual low competition (top) and high competition (bottom) social events from the Multisensory Attention Assessment Protocol.

Figure 2. Static image of the dynamic audiovisual social events from the Intersensory Processing Efficiency Protocol.

III. MANUSCRIPT TWO:

INTERSENSORY PROCESSING OF SOCIAL EVENTS AT 6 MONTHS PREDICTS LANGUAGE OUTCOMES AT 18, 24, AND 36 MONTHS OF AGE

Abstract

Intersensory processing of social events (e.g., matching sights and sounds of audiovisual speech) is a critical foundation for language development. Two recently developed protocols, the Multisensory Attention Assessment Protocol (MAAP) and the Intersensory Processing Efficiency Protocol (IPEP), assess individual differences in attention and intersensory processing at a sufficiently fine-grained level for predicting developmental outcomes. Recent research using the MAAP demonstrates that 12-month intersensory processing of face-voice synchrony predicts language outcomes at 18 and 24 months, holding traditional predictors (parent language input, SES) constant. The present study builds on these findings testing younger infants using the IPEP, a more comprehensive, fine-grained index of intersensory processing. Using a longitudinal sample of 103 infants, we tested whether intersensory processing (speed, accuracy) of faces and voices at 6 months would predict language outcomes at 18, 24, and 36 months, holding traditional predictors constant. Results demonstrate that intersensory processing of faces and voices at 6 months accounted for significant unique variance in language outcomes beyond that of traditional predictors. Findings indicate that intersensory processing of faces and voices is an important foundation for language development, can be assessed at the individual participant level by 6 months, and predicts language outcomes even 2.5 years later.

Keywords: intersensory processing, individual difference measures, parent language input, child language, audiovisual speech perception

Introduction

Parent language input is a well-established predictor of child language development (Hart & Risley, 1995; Hoff, 2003; Huttenlocher et al., 1991; Rowe, 2008). Greater quantity (amount) and quality (diversity) of parent language input are associated with better child language outcomes (Hart & Risley, 1992; Huttenlocher et al., 1991; Weisleder & Fernald, 2013; Weizman & Snow, 2001). In contrast, there has been little research characterizing the role of individual differences in intersensory processing of audiovisual events (e.g., audiovisual speech) as a foundation of child language outcomes, despite agreement that it is an important early foundation for language development (Bahrick et al., 2020; Bahrick, Todd et al., 2018; Edgar et al., under review). Two recently developed measures, the Multisensory Attention Assessment Protocol (MAAP, Bahrick, Todd et al., 2018) and the Intersensory Processing Efficiency Protocol (IPEP; Bahrick, Soska, et al., 2018) now allow researchers to assess fine-grained individual differences in multisensory attention and intersensory processing in young infants in a context highly relevant for language acquisition, that of dynamic faces and voices during audiovisual speech.

The MAAP assesses individual differences in three "multisensory attention skills"—sustaining attention, shifting/disengaging attention, and intersensory processing (matching synchronous sights and sounds)—for both audiovisual social (speech) and nonsocial (object) events. Using this measure, we recently found that intersensory processing (but not sustaining or shifting/disengaging attention) for social (but not nonsocial) events at 12 months of age was a strong predictor of child language outcomes. It predicted child speech production and expressive vocabulary at 18 and 24 months of

age, even after controlling for two other well-established predictors: parent language input, and SES (Edgar et al., under review). These findings replicate and extend our prior research with toddlers and young children (2- to 5-year-olds) demonstrating that intersensory processing of faces and voices predicts receptive and expressive language outcomes (Bahrick, Todd, et al., 2018). Together, these findings indicate that intersensory processing of faces and voices in infancy and early childhood provides an important foundation for language development. Further, by assessing individual differences in intersensory processing we can predict which children will benefit most from the parent language input and other language learning opportunities provided by their environment.

The present study builds on the study by Edgar et al. (under review), extending it to younger infants (6-month-olds), later language outcomes (18, 24, and 36 months of age), and a different measure of intersensory processing (IPEP). Given that our prior study demonstrated that intersensory processing (but not attention maintenance or shifting/disengaging speed) was a strong predicter of language outcomes, in this study, we focused specifically on intersensory processing. The IPEP is an ideal measure for this purpose. It is a fine-grained measure of just intersensory processing skills, assessing accuracy as well as speed of matching sights and sounds. Here, we demonstrate the viability of the IPEP as an index of individual differences in intersensory processing for evaluating developmental relations with language outcomes. Unlike the MAAP, which indexes accuracy of intersensory matching using a traditional two-screen intermodal matching method, the IPEP indexes both accuracy and speed of intersensory matching of a single synchronous target event in the presence of five asynchronous distractor events. Like the MAAP, it assesses intersensory matching for both social and nonsocial events.

However, given that audiovisual speech events provide the most relevant context for language learning and that both our prior studies (Bahrick, Todd et al., 2018; Edgar et al., under review) found that attention to social, but not to nonsocial events, predicted language outcomes, in this study we again focused on intersensory processing of faces and voices rather than object events.

Finally, the present study also extends our prior findings by investigating individual differences in intersensory matching speed and accuracy in younger infants of 6 months. Given that intersensory processing skills develop rapidly across the first 6 months of life (for reviews, see Bahrick et al., 2020; Bremner et al., 2012) might they also predict child language outcomes at 18 and 24 months? Further, we also tested if they would predict child language outcomes even later in development, at 36 months of age. Similar to our prior study, we investigated to what extent individual differences in intersensory processing of faces and voices during natural, synchronous, audiovisual speech would predict language outcomes given comparable levels of parent language input and SES. We expected that findings would parallel and extend those of our prior study and indicate that infants with better intersensory processing of faces and voices at 6 months of age would show greater language outcomes than those with poorer intersensory processing skills. This would suggest that by 6 months of age, individual differences in intersensory processing are meaningful predictors of later language and that infants who show greater intersensory matching skills are able to benefit more from language learning opportunities provided by their language environment.

Intersensory Processing of Audiovisual Speech: A Foundation for Child Language Development

Intersensory processing is a fundamental basis for guiding infant selective attention and perceptual development (Bahrick et al., 2020; Bahrick & Lickliter, 2012; E. J. Gibson, 1969). It helps infants direct attention to multimodal stimulation provided by unitary events (e.g., faces and voices of a person speaking) and filter out irrelevant stimulation form co-occurring events (e.g., a nearby conversation, or activity). A skill that develops in early infancy, intersensory processing involves detecting intersensory redundancy (the synchronous co-occurrence of stimulation across two or more senses). Intersensory redundancy (when the same information is presented simultaneously and synchronously across the senses) is highly salient and recruits attention to properties of events that are common across sense modalities (i.e., amodal information) including temporal synchrony, rhythm, intensity, and tempo. Most events provide multiple forms of amodal information (Bahrick, 2010; Bahrick et al., 2020; Bahrick & Todd, 2012). One property of amodal events, temporal synchrony (simultaneous changes in patterns of visual and acoustic stimulation, including auditory and visual onset, offset, duration, and common temporal patterning), is proposed to be the 'glue' that binds stimulation across the senses (Bahrick & Lickliter, 2002; Lewkowicz, 2000b) is considered a global amodal property that facilitates the detection of other (nested) amodal properties including duration, rhythm, and tempo (Bahrick, 1992, 1994, 2001). Here, we focus on intersensory processing across the auditory and visual modalities.

In face-to-face interactions, the perceiver can both hear what is said and see the corresponding articulatory gestures (Kuhl & Meltzoff, 1982; Rosenblum, 2008;

Stevenson et al., 2014). When a person speaks, they provide highly salient amodal information—their vocalizations and mouth movements are spatially co-located, and share common rhythm, tempo, and intensity shifts (Gogate et al., 2001; Gogate & Hollich, 2010). Further, mapping a word onto an object is a multisensory activity involving linking a sound with a visual object or event. Parents intuitively use intersensory redundancy to help infants learn language, often labeling an object while holding and moving it in synchrony with its label (Gogate et al., 2000). The shared onset, offset, and duration of the simultaneous movement and naming recruits selective attention and provides salient amodal information that links the speech sounds with the object. Although the relation between an object and its name is arbitrary to a word-mapping novice, the intersensory redundancy provided by the simultaneous movement and labeling reduces the uncertainty about the word-referent relation (Gogate & Hollich, 2010). We have proposed that better intersensory processing skills promote more accurate and efficient processing of audiovisual speech events, allowing infants to take greater advantage of parent language input and language learning opportunities such as word mapping (Bahrick et al., 2020; Edgar et al., under review).

Infant intersensory processing has been studied extensively with methods appropriate for group-level analyses including the intermodal preference (Bahrick, 1983, 1988; Lewkowicz, 1992; Spelke, 1976) and habituation (Bahrick & Lickliter, 2004; Bahrick & Pickens, 1988; Caron et al., 1988; Lewkowicz, 2000, 2003; Walker-Andrews & Grolnick, 1983). These methods have been used to assess intersensory processing skills for groups of infants at specific ages. Studies using these methods have revealed that infants can match faces and voices on the basis of a wide range of amodal properties.

For example, newborns can detect face-voice synchrony in point-light displays of a woman speaking (Guellaï et al., 2016) and in nonhuman primate species on the basis of temporal synchrony (Lewkowicz et al., 2010). Infants from 2 to 4 months of age can match vowel sounds with the corresponding shape of lip movements on the basis of spectral information in the vowel sounds (Kuhl & Meltzoff, 1982; Patterson & Werker, 1999, 2003). By 4 to 7 months of age infants match faces and voices on the basis of the speaker's affect (Soken & Pick, 1992; Vaillant-Molina et al., 2013; Walker, 1982), gender (Richoz et al., 2017; Walker-Andrews et al., 1991), and age (Bahrick et al., 1998). Findings also demonstrate that infants can perceive amodal properties provided by nonsocial events such as objects dropping or striking a surface, including temporal synchrony, rhythm, tempo, and temporal microstructure, including object substance and composition (see Bahrick, 1983, 1987, 1988; Bahrick et al., 2002; Bahrick & Lickliter, 2000, 2004; Lewkowicz, 1992; Lewkowicz & Marcovitch, 2006). Thus, group-level studies assessing accuracy of intersensory processing reveal that infants can detect audiovisual synchrony and match faces and voices under a variety of conditions across the first half year of life. In contrast, speed of intersensory processing (how quickly infants find the synchronous audiovisual event) has received virtually no research focus in infants, although it has been assessed in adults (Fiebelkorn et al., 2012).

Studies using these methods designed for group-level analyses have indicated that intersensory processing in infancy serves as a foundation for language development. For example, studies have demonstrated that synchronous, but not asynchronous, object movement and labeling promotes object-label matching in infants and toddlers (Gogate et al., 2006; Gogate & Bahrick, 1998; Jesse & Johnson, 2016). These methods, however, are

not designed to provide scores for individual infants and are thus not appropriate for predicting outcomes or assessing change across development.

In contrast, individual difference measures assessing the skills of individual infants relative to one another can address the extent to which intersensory processing in infancy predicts individual differences in outcomes such as language, social, or cognitive functioning. This was explored in two recent studies using the MAAP. Intersensory processing of face-voice synchrony (a woman telling a story) in 2- to 5-year-old children assessed by the MAAP was found to predict both receptive and expressive language (Bahrick, Todd et al., 2018). In another study testing younger children (Edgar et al., under review), we found that intersensory processing of face-voice synchrony on the MAAP at 12 months predicted child quality and quantity of speech production at 18 and 24 months, as well as expressive vocabulary size at 18 and 24 months. Moreover, intersensory processing at 12 months predicted a large, significant amount of unique variance in language outcomes, over and above that of well-established predictors including parent language input (quality and quantity) and SES. Thus, 12 months was found to be an important time in development for investigating the role of intersensory processing skills in predicting outcomes. This is, in part because these skills are still undergoing significant development across the first year of life. However, even in the first 6 months of life, infants learn to efficiently locate the source of a sound, in both social and nonsocial events, while filtering out other concurrent auditory and visual stimulation (e.g., Bahrick, 1983; Kuhl & Meltzoff, 1982; D. J. Lewkowicz, 1992; Spelke, 1976). Might intersensory processing as early as 6 months of age, also predict language

outcomes even in the context of other well-established predictors? We addressed this question using the IPEP.

The Intersensory Processing Efficiency Protocol (IPEP)

The Intersensory Processing Efficiency Protocol (IPEP; Bahrick, Soska et al., 2018) provides a fine-grained measure of just intersensory processing, including both speed and accuracy of detecting audiovisual synchrony in social and nonsocial events. The IPEP features six concurrent, dynamic events (both social and nonsocial conditions), and infants must detect the sound-synchronous target event from among five competing visual distractor events that are asynchronous with the soundtrack. It simulates the "noisiness" of the natural environment, resembling the task of picking out a speaker from a crowd. The proportion of total looking time to the sound synchronous event serves as an index of the accuracy of intersensory matching (similar to the intermodal preference method) and the latency to shift attention to the sound-synchronous event is an index of the speed of intersensory processing (i.e., speed of selecting the audiovisual synchronous target). Individual scores are derived for each measure across a number of trials for each infant, making it fine-grained enough to reliably predict outcomes, with a relatively stable mean. Finally, the IPEP does not require verbal responses or language comprehension, making it appropriate for use with preverbal infants and children.

Well-Established Predictors of Child Language: Parent Language Input and SES

Parent language input is a well-established predictor of child language outcomes. Both the quantity and quality of parent language input are positively related to child language outcomes (Hart & Risley, 1995; Huttenlocher et al., 1991; Rowe, 2012). Parents who speak more words and provide higher quality language input provide more

opportunities for children to hear, and in turn learn, new words. Parent language input has been conceptualized a variety of ways. For the purposes of the present study, we refer to quantity as the total number of words spoken to the child (word tokens; (Jones & Rowland, 2017; Rowe, 2012; Soderstrom et al., 2018; Weisleder & Fernald, 2013), and we refer to quality as lexical diversity, the number of different words spoken to the child (word types; Malvern & Richards, 2012; McCarthy & Jarvis, 2010). Using these measures of quality (types) and quantity (tokens), research has demonstrated that parent language input after 12 months is a strong predictor of child language outcomes (Edgar et al., under review; Gilkerson et al., 2018; Jones & Rowland, 2017; Pan et al., 2005). In contrast, few studies have assessed quality and quanity of parent language input earlier than 12 months of age as predictors of child language outcomes. In our prior study using the MAAP, we found that while controlling for SES and intersensory processing at 12 months, quality and quantity of parent language input at 12, 18, and 24 months predicted child language outcomes at 18 and 24 months of age. However, of the three ages, parent language input at 12 months was the weakest predictor of child language outcomes at 18 and 24 months (Edgar et al., under review). Thus, it is unclear if quality and quantity of parent language input at 6 months will be strong predictors of child language outcomes at 18, 24, and 36 months.

Socioeconomic status (SES) is also a well-known predictor of child language, with higher SES predicting greater quality of parent language input (Hart & Risley, 1995; Rowe, 2018), which in turn predicts increases in later child vocabulary (Hoff, 2003). Parents with more education use a greater number of unique words and complex

utterances when speaking to their children than parents with less education (Bornstein et al., 1998; Rowe et al., 2005).

A number of other predictors of child language outcomes related to intersensory processing have been studied but were not examined in the present study. One well-known predictor of language outcomes in toddlers is speech processing efficiency, the ability to quickly and accurately link spoken words with their referents (Fernald et al., 2006; Marchman & Fernald, 2008). Speech processing efficiency speed and accuracy have been found to predict vocabulary growth across the second year of life (Fernald et al., 2006). Here, we focus on intersensory processing skills given they are earlier developing skills that potentially cascade into later developing skills including speech processing efficiency, joint attention, word mapping, and fluency and connectedness of interaction.

The Present Study

The present study uses the IPEP to examine the unique contribution of intersensory processing at 6 months in predicting child language outcomes at 18, 24, and 36 months. We expect results to parallel and extend those of our previous study (Edgar et al., under review) assessing intersensory processing of social events with the MAAP. That is, we expect that intersensory processing (measures of speed and accuracy) of social events at 6 months of age will predict significant unique variance in child language outcomes at 18, 24, and 36 months, while controlling for parent language input and SES.

Method

Participants

One-hundred and four infants participated as part of a larger ongoing longitudinal study on the development of multisensory attention skills and language, cognitive, and social outcomes. The ongoing longitudinal study, entitled "[blinded]", received IRB approval from the Social and Behavioral Review Board of [blinded IRB #]. The final sample consisted of a total of $N = 103$ infants (one infant participated at 6 months and none of the other ages, and thus was excluded from analyses). Infants were assessed at 6, 18, 24, and 36 months. Demographic information for the sample can be found in Table 1. For a summary of the assessments administered at each age and dependent variables, see Table 2.

Child Intersensory Processing Measures: IPEP

Stimulus Events

The IPEP consists of 48 8-s trials with 24 social and 24 nonsocial trials presented in four alternating blocks of 12 trials each. This updated version of the IPEP was modified and refined based on stimuli and procedures used in Bahrick, Soska et al. (2018) including filming new social events (see Figure 1), increasing trial length from 6- to 8-s to be more appropriate for younger infants, and making social/nonsocial trial blocks a within participants factor. As before, trials consisted of a 2 (rows) x 3 (columns) grid of 6 dynamic social or nonsocial events. The entire grid was 67.3 x 38.1 cm (51.3 degrees visual angle), and each square of the grid covered 20.3 x 16.5 cm (16.5 degrees visual angle). The social events depict six women, each telling a different story using

infant-directed speech. Nonsocial events depict six wooden objects (single objects or clusters of objects) being dropped on a surface in erratic temporal patterns. On each trial, the natural soundtrack is synchronized with the movements of one event while the movements of the other five events are asynchronous with the soundtrack. Thus, the infant's task is to visually fixate the sound-synchronous speaker (target event) amidst the five asynchronous distractors on each trial. For an example video, please visit https://nyu.databrary.org/volume/336. A smiley face is presented zooming in and out for two-seconds between each trial to attract the infant's attention to the center of the screen. Six different types of smiley faces, each of different primary colors, were presented in a pseudorandom order across trials.

Procedure

A 119.4-centimeter widescreen monitor (NEC Multisync PV61) was used to present the IPEP and a Tobii X120 eye-tracker was used to record gaze fixations. Infants were seated on their caregiver's lap approximately 70 centimeters in front of the monitor, and 60 centimeters in front of the Tobii eye-tracker. The eye-tracker, located directly under the monitor, was tilted upward, 20 degrees, towards the child's eyes. An experimenter, seated behind the child, presented the stimulus events to the monitor using Tobii Studio (Version 3) from a computer (Mac Pro Computer with 16 GB of RAM, a 3.33-GHz processor, and a 400-MHz graphics card). Caregivers wore black-out glasses to ensure they were unaware of the location of the sound-synchronous target event.

The experimenter viewed a live recording from a video camera (SONY FDR-AX33) placed facing the infant to ensure the infant was seated in an optimal position for eye-tracking calibration and for viewing the stimuli. Tobii Studio's "Infant" 5-point

calibration procedure was used to calibrate the infrared corneal reflection-to-pupil tracking system for each infant. The experimenter calibrated the infant's eye-gaze to five points on the widescreen monitor for accurate calculation of infant visual fixations during the procedure.

The 24 social and 24 nonsocial trials were arranged into four alternating blocks of 12 trials each (social, nonsocial, social, nonsocial, or vice versa, counterbalanced across participants). Social and nonsocial events in the IPEP are presented in separate blocks and designed to be analyzed separately depending on the research focus of the study. The present study focuses on social events, given the importance of audiovisual speech perception for predicting language outcomes, as well as results of our previous studies using the MAAP in which performance on social (but not nonsocial) trials predicted language outcomes (Edgar et al., under review).

Infant eye gaze was sampled at 120Hz by the Tobii X120 system. The number of usable trials ranged from 21 to 48 trials, with an average of 43.33 (*SD* = 6.08 trials) out of 48 trials. Trials in which infants were inattentive (less than 250 ms looking to the screen) were excluded from analyses. Further details regarding eye-tracking parameters and data processing are presented in the Supplement (p. 1).

IPEP Measures

The IPEP provides three measures of intersensory processing: accuracy of intersensory matching, speed of intersensory matching, and frequency of intersensory matching. In the present study, we focus on just two of these measures: accuracy of intersensory matching (duration of looking) and speed intersensory matching (reaction time to fixate). Frequency of intersensory matching (proportion of total trials on which

the infant fixated the sound synchronous target event) was not significantly correlated with any of our outcome variables and was thus excluded from subsequent analyses. Accuracy of intersensory matching (PTLT; proportion of total looking time to the sound-synchronous "target" event) is the traditional measure used in studies of intersensory processing and it assesses how long the infant fixates the sound-synchronous visual event. Greater looking to the sound-synchronous event provides an opportunity for longer and deeper processing of the multimodal event. PTLT was calculated by dividing the looking time to the AOI depicting the sound-synchronous target event by the total looking time to all five AOIs depicting sound-asynchronous "distractor" events. PTLTs greater than .167 (chance) indicate a preference for the sound-synchronous target event. Speed of intersensory matching (RT) assesses how quickly infants visually fixate the sound-synchronous event. Faster speeds in fixating the target event reflect faster intersensory matching and more time for processing the multimodal event. RT was calculated as the latency from trial onset to produce a fixation (of at least 50-ms) to the sound-synchronous event. PTLTs and latency scores were calculated on each trial was then averaged across all trials within each condition (social, nonsocial).

Parent Language Input and Child Language Production Measures

Parents and children participated in an 8-minute (M = 8.15 minutes, range = 3.30 to 12.28) semi-structured Parent Child Interaction (PCI). In a lab playroom, the parent and child were seated facing each other at a table (40 X 28 in.) in the center of the room (see Figure 2). At 6, 12, and 18 months, children sat in a seat attached to the edge of the table, and a 24 and 36 months, they sat in a booster seat attached to a chair.

At each age, parent and child speech during the PCI was transcribed by trained research assistants who watched the video recordings. Transcription units were words. A second trained research assistant checked the original transcriptions to establish reliability. Any disagreements between the primary transcriber and the secondary transcriber were decided by a third research assistant, who was not aware of the topic of disagreement. The Child Language Data Exchange Systems (CHILDES; MacWhinney, 2000) FREQ program was used to calculate the quantity (tokens; total number of words spoken) and quality (types; total number of different, or unique, words spoken) of parent language input and child language production. To equate across PCIs of different durations, a per-minute ratio was calculated by dividing the number of types (or tokens) by the duration of the interaction. Only speech directed to the child was included in type and token calculations (e.g., parents rarely spoke to the experimenter, but this speech was not transcribed).

Child Vocabulary Measures

At 18 and 24 months, parents completed the Mac-Arthur Bates Communicative Development Inventory (MB-CDI) in either English (Fenson et al., 2007) or Spanish (Jackson-Maldonado et al., 2003) or both, depending on parental report of the child's primary language (for details, see Supplement, pp. 1-2). At 36 months, children received the Peabody Picture Vocabulary Test – 4th Edition (PPVT; Dunn & Dunn, 2007) to assess the child's receptive vocabulary size and the Expressive Vocabulary Test – 2nd Edition (EVT; Williams, 2007) to assess the child's expressive vocabulary size.

Results

Data Analysis Overview

The present study examined the extent to which intersensory matching (both speed and accuracy) of social events at 6 months predicted child language outcomes at 18, 24, and 36 months, while holding constant other well-known predictors of child language, including parent language input at 6 months (both quantity and quality) and SES. We first conducted correlations between the 6-month predictors (speed and accuracy of intersensory matching, quantity and quality of parent language input, and maternal education) and child language outcomes (child quantity and quality of speech and expressive vocabulary at 18, 24, and 36 months and receptive vocabulary at 18 and 36 months). Our primary analyses designed to address our research questions consisted of multiple regressions with five predictors of language outcomes: accuracy of intersensory matching, speed of intersensory matching, quality (types) of parent language input, quantity (tokens) of parent language input, and maternal education. With a sample size of $N = 103$, there is sufficient power for multiple regression analyses to detect a non-zero path coefficient that accounts for 6% unique variance (assuming a ß of .80, a two-tailed p-value of .05, five predictors, and an R^2 of .30).

Robust Full Information Maximum Likelihood (FIML) estimation was used for all analyses in MPlus (Version 1.6). Missing data ranged from 6.7% (maternal education) to 51% (CDI receptive and expressive vocabulary; see Table 3). To ensure that data were not systematically missing or missing not at random (MNAR or non-ignorable missingness; Rubin, 1976), we conducted missing value analyses using various techniques (e.g., *t*-tests, logistic regression, Little's MCAR test). T-tests and logistic

regressions revealed that missingness was not related to any of the main predictors or outcomes. From these analyses, we concluded that the data were missing at random (MAR; Rubin, 1976), supporting the use of FIML.

Secondary analyses were conducted to assess the influence of language spoken at home, gender, race, and ethnicity as covariates in predicting child language outcomes. Overall, their inclusion did not change the strength of the main predictors in predicting the child language outcome measure (for details, see Supplement, pp. 2-5). Thus, the present study did not include home language, gender, race, or ethnicity as covariates in the main analyses.

The present study focused on 6-month intersensory matching (both speed and accuracy) of the social (audiovisual speech) events, an important language learning context for children. However, we also conducted supplemental analyses of 6-month intersensory matching (both speed and accuracy) of nonsocial events as predictors of child language outcomes at 18, 24, and 36 months (see Supplement pp. 5-7 and Supplemental Tables 1-2 for details). Speed of intersensory matching (but not accuracy) for nonsocial events predicted some language outcomes (quantity of child speech, receptive and expressive vocabularies), at one of the ages, 36 months (but not 18 or 24 months) after controlling for quantity and quality of parent language input at 6 months, and maternal education.

Correlational Analyses

Descriptive statistics for 6-month intersensory matching (speed and accuracy) of social events, 6-month parent language input (quantity and quality), and child language outcomes at 18, 24, and 36 months are displayed in Table 3 and correlations among these

variables are displayed in Table 4. We first calculated Pearson correlation coefficients using FIML[1] and correcting for familywise error rate[2]. Correlations were conducted between our main predictor variables—accuracy of intersensory matching for social events at 6 months, speed of intersensory matching at 6 months, quality of parent language input at 6 months, quantity of parent language input at 6 months, and maternal education—and our language outcome variables—quality of child speech production, quantity of child speech production, receptive vocabulary, and expressive vocabulary—at 18, 24, and 36 months of age (there was no measure of receptive vocabulary size at 24 months). Several novel findings emerged.

Results of our correlational analyses (see Table 4) revealed that accuracy of intersensory matching of social events at 6 months was a remarkably strong predictor of language outcomes at 18, 24, and 36 months of age (*r*-range: .25-.40, *p*s < .01), with greater intersensory matching of social events predicting greater quality and quantity of child speech production and larger expressive vocabulary size. Overall, greater quality and quantity of parent language input at 6 months (*r*-range: .21-.28, *p*s < .01), as well as greater maternal education (*r*-range: .23-.46, *p*s < .01), predicted better child language outcomes, particularly at 24 and 36 months, consistent with prior findings. These

[1] All correlations conducted using FIML were compared to traditional bivariate pairwise Pearson's *r* correlations (excluding participants with missing data) to ensure that findings derived from FIML were similar to the general pattern of findings from participants with complete data. All findings using FIML paralleled those of the bivariate pairwise correlations, with similar magnitudes and directions.

[2] At 18 and 36 months, there were four child language outcomes (child speech production: quantity and quality; child vocabulary size: receptive and expressive) and thus we used a familywise significance level of $p < 0.0125$ (.05 / 4; two-tailed) to evaluate results. At 24 months, there were three child language outcomes (child speech production: quantity and quality; child expressive vocabulary size) and thus we used a familywise significance level of $p < .0167$ (.05 / 3; two-tailed) to evaluate results.

correlational analyses informed our multiple regression models (see Supplement, pp. 7-8, for more details.)

Multiple Regression Analyses: Intersensory Matching Predicts Language Outcomes

Our primary analyses consisted of multiple regressions assessing the role of intersensory processing of social events on language outcomes in the context of other predictors. Findings from our correlational analyses revealed that accuracy of intersensory matching of social events at 6 months predicted a variety of child language outcomes at 18, 24, and 36 months. However, these findings do not reveal relative importance or aggregate effects of multiple predictors, including SES, parent language input, and intersensory matching on language outcomes. Might accuracy of intersensory matching of social events at 6 months remain a significant predictor of child language outcomes when holding constant parent language input and SES? What is the relative predictive power (unique variance) of each of these variables in predicting language outcomes when the others are controlled? What are the aggregate effects (total variance accounted for) of all of these variables together in predicting language outcomes? To address these key research questions, our main analyses consisted of multiple regressions conducted using FIML to assess the role of 6-month accuracy of intersensory matching of social events, SES, and parent language input as a predictors of child language outcomes. We conducted regression analyses for each of the 11 child language outcomes including quality of child speech (18, 24, and 36 months), quantity of child speech (18, 24, and 36 months), receptive vocabulary (18 and 36 months; there was no measure of receptive vocabulary at 24 months), and expressive vocabulary (18, 24, and 36 months). We also included speed of intersensory matching in our models given that it has typically not been

assessed along with accuracy of matching nor have studies previously assessed it as a predictor of language outcomes. We controlled for both quantity and quality of parent language input at 6 months and maternal education to examine the extent to which intersensory matching of social events predicted language outcomes at 18, 24, and 36 months, over and above that of these well-established predictors at 6-months of age. Secondary analyses also assessed the role of parent language input at older ages, given it was found to be a stronger predictor later in development (Edgar et al., under review; see Supplement, pp. 12-15 and Supplemental Tables 10 through 16).

For each of the 11 outcome variables, we conducted five multiple regression models to assess the amount of unique variance (ΔR^2) attributable to each predictor in predicting the outcome variable. The unique variance attributable to a given predictor is the change in R^2 when that predictor is entered last in the regression model (i.e., holding all other predictors constant). To accomplish this, each of the five predictors at 6 months (accuracy of intersensory matching, speed of intersensory matching, parent speech quality, parent speech quantity, and maternal education) was entered into the regression model in a different order (1^{st}, 2^{nd}, 3^{rd}, 4^{th}, 5^{th}). For example, in Model 1 we derived the unique variance attributable to accuracy of intersensory matching in predicting an outcome by entering it last (i.e., holding constant all other predictors entered earlier: maternal education, quality of parent language input, quantity of parent language input, speed of intersensory matching, and so forth for Models 2 through 5; for details, see Supplemental Tables 4 through 9). The amount of total variance explained by all 5 predictors, as well as the unique variance explained by each predictor in predicting each of the language outcomes at each age are summarized in Table 5.

Overall, the five predictors accounted for a significant amount of total variance in 5 of the 11 child language outcomes: child speech quality at 18 months, child speech quality and quantity at 24 months, and receptive and expressive vocabulary at 36 months (range: 27% – 35%, *p*s < .01; see Table 5). Remarkably, 6-month accuracy of intersensory matching of social events was a significant predictor and accounted for a large amount of unique variance in these 5 child language outcomes (range: 8% to 15%, *p*s < .05), as well as significant unique variance in 3 other child language outcomes (child speech quantity at 18 months, and expressive vocabulary at 18 and 24 months; range: 6% to 11%, *p*s < .05). In contrast, 6-month speed of intersensory matching accounted for a smaller but significant amount of unique variance (range: 3% to 4%, *p*s < .05) in child speech quality and quantity at 18 months (but not at 24 or 36 months). Further, maternal education accounted for a significant amount of unique variance in 6 of the 11 child language outcomes, child speech quality at 18, 24, and 36 months, child speech quantity at 24 months, and expressive and receptive vocabulary at 36 months (range: 6% to 17%, *p*s < .05). Six-month parent language input (both quantity and quality) accounted for a non-significant amount of unique variance in most child language outcomes (range: 0 – 4%, *p*s > .05) with just one exception: 6-month parent language quality significantly predicted receptive vocabulary at 18 months, $p < .05$. Details regarding the amount of unique variance attributed to each child language outcome by each predictor, as well details quantifying relations between 6-month accuracy and speed of intersensory matching of social events and child language outcomes can be found in the Supplement, pp. 8-12 and Supplemental Tables 4 through 9.

Supplemental Analyses: Parent Language Input at Older Ages Predicts Language Outcomes

Parent language input (quantity and quality) at 6 months was a weak predictor of child language outcomes at 18, 24, and 36 months, controlling for speed and accuracy of intersensory matching at 6 months and maternal education. However, parent language input at older ages—18, 24, and 36 months—was a moderately strong predictor of child language outcomes (for details on correlational analyses, see Supplemental Table 3). When parent language input (quality and quantity) at older ages was substituted for 6-month parent language input in our multiple regression models, results indicated that parent language input at 24 months significantly predicted child language at 24 months, and parent language input at 36 months significantly predicted child language at 36 months, *p*s < .05, holding all other predictors constant. In contrast, parent language input at 18 months was not a predictor of child language at 18 months. For details on these secondary regression analyses, see Supplementary Material, pp. 12-15.

Importantly, accuracy of intersensory matching of social events at 6 months remained a strong predictor of language outcomes, even after holding quantity and quality of parent language input at 18, 24 and 36 months constant. Thus, when children receive equal amounts of parent language input at 6, 18, 24, and 36 months, accuracy of intersensory matching of faces and voices at 6 months continues to explain a significant proportion of leftover variability in child language outcomes.

Summary: Unique Contributions of Intersensory Processing to Child Language Outcomes

In sum, multiple regression analyses indicated that intersensory matching, particularly accuracy of matching social events (faces and voices) at 6 months was a significant predictor of child language outcomes at 18, 24, and 36 months, even after controlling for a variety of parent variables (parent quality and quantity of language, maternal education). Most notably, accuracy of matching faces and voices at 6 months predicted significant unique variance in child language outcomes at 18, 24, and 36 months, even when parent variables (quality and quantity of parent language input, maternal education) were held constant. In particular, it significantly predicted unique variance in expressive vocabulary at all 3 ages, as well as quality and quantity of child speech production at 18 and 24 months. Thus, when infants receive equal amounts of quantity and quality of parent language input and have parents with similar levels of maternal education, 6-month-olds with greater accuracy of intersensory matching of faces and voices show better language outcomes at 18, 24, and 36 months. In contrast, speed of intersensory matching and parent language input (quality and quantity) at 6 months rarely predicted significant unique variance in any child language outcomes. Finally, maternal education predicted a variety of child language outcomes at 18, 24, and 36 months, indicating a language advantage for children of mothers with more education.

Discussion

In the present study, we examined the contribution of accuracy and speed of intersensory processing of social events (faces and voices) at 6 months of age as a

predictor of child language outcomes at 18, 24, and 36 months, along with well-established predictors including parent language input (quantity and quality) at 6 months and SES (maternal education). Results revealed that the accuracy of intersensory matching for social events at 6 months predicted child language outcomes at 18, 24, and 36 months, over and above the contribution of parent language input and SES. These results indicate that the accuracy of intersensory processing skills at 6 months (maintaining attention to the face of a person speaking amidst other dynamic faces) is a strong and independent predictor of child language development across the second and third years of life. These findings are elaborated below.

Intersensory Matching of Social Events at 6 Months Predicts Multiple Child Language Outcomes at 18, 24, and 36 Months

Our prior findings demonstrated that intersensory processing of faces and voices at 12 months predicted language outcomes at 18 and 24 months, even when controlling for parent language input (quality and quantity) and SES (maternal education; Edgar et al., under review). In the present manuscript, our main research question focused on whether and to what extent accuracy and speed of intersensory matching of faces and voices at 6 months would predict child language outcomes at 18, 24, and 36 months when controlling for concurrent parent language input and SES. Overall, our findings converge with our prior findings that intersensory processing of faces and voices in infancy predicts language outcomes in later development, and they also extend our original findings in important ways.

In the present study, we found that accuracy of intersensory matching of faces and voices at 6 months predicted vocabulary across the first three years of life as well as

child speech production across the first two years of life. Specifically, accuracy of matching faces and voices in infancy was a strong and independent predictor vocabulary size in toddlerhood, including expressive vocabulary at 18, 24, and 36 months (predicting 6% to 8% unique variance), and receptive vocabulary at 36 months (predicting an impressive 15% unique variance). It was also a strong and independent predictor of child speech production in toddlerhood, including quality and quantity of child speech at 18 and 24 months (predicting 10% to 14% unique variance). Critically, it predicted child language outcomes holding traditional predictors such as parent language input and SES constant. These findings parallel our prior findings that intersensory processing of faces and voices in later infancy (12 months) predict child language outcomes in toddlerhood (18, 24 months; Edgar et al., under review). However, they also extend our prior findings, demonstrating that intersensory processing in early infancy, at 6 months of age, predicts child language outcomes not only at 18 and 24 months, but also at 36 months of age. These novel findings indicate that at 6 months, given equal amounts of parent language input (quantity and quality) and SES, the accuracy of intersensory processing of faces and voices can predict which children will benefit the most from language learning opportunities provided by parent language input. Infants who show greater accuracy of intersensory processing at 6 months of age go on to show greater language outcomes a year later at 18 months, a year and a half later at 24 months, and two and a half years later at 36 months of age.

In contrast, speed of intersensory matching of faces and voices at 6 months was a weaker predictor of child language outcomes than accuracy of matching, predicting only child speech production (quantity and quality) at 18 months of age over and above other

predictors (predicting 3% to 4% unique variance). Thus, at 6 months of age, although speed of intersensory matching predicts some outcomes at 18 months, accuracy of intersensory matching appears to be a much stronger predictor of multiple child language outcomes (both child vocabulary and speech production) and predicts outcomes later in development (24 and 36 months) than speed of intersensory matching.

Overall, convergent findings across two different protocols, the MAAP and IPEP (Bahrick, Todd et al., 2018; Edgar et al., under review), demonstrate that individual differences in intersensory processing of faces and voices in the first year of life are meaningful predictors of later language outcomes while holding constant traditional predictors of parent language input and SES. Infants who show greater intersensory matching skills appear to benefit more from language learning opportunities provided by their language environment and go on to show greater vocabulary and speech production skills across the next two and a half years of life.

Intersensory Matching of Social Events at 6 Months Predicts Child Language Outcomes, Controlling for Parent Language Input at Older Ages (18, 24, 36 months)

Accuracy of intersensory matching of faces and voices at 6 months was a strong predictor of multiple language outcomes at 18, 24, and 36 months, holding parent language input (quality and quantity) at 6 months constant. Further, parent language input at older ages (18, 24, 36 months) was a moderately strong predictor of child language outcomes (see Supplemental Analyses: Parent Language Input at Older Ages Predicts Language Outcomes). However, our supplementary analyses demonstrated that intersensory matching of social events at 6 months remained a strong predictor of child language outcomes even when holding parent language input at 18, 24, and 36 months

constant. After controlling for parent language input at older ages, accuracy of intersensory matching of faces and voices at 6 months predicted expressive vocabulary size at 18, 24, and 36 months (predicting 5% to 13% unique variance), as well as child speech production (quality and quantity) at 18 and 24 months (predicting 12% to 17% unique variance; for summary of the unique variance explained by each predictor in predicting each of the language outcomes at each age, see Supplemental Table 16). Thus, given equal amounts of parent language input at 18, 24, and 36 months, accuracy of intersensory matching of faces and voices at 6 months still predicts which children will benefit the most from parent language input later in development.

SES (Maternal Education) Also Predicts Multiple Child Language Outcomes at 18, 24, and 36 Months

Maternal education, an index of SES, was also a strong and significant predictor of multiple child language outcomes, holding constant accuracy and speed of intersensory matching and parent language input (quality and quantity). It especially predicted child language outcomes at older ages, particularly expressive and receptive vocabulary at 36 months (predicting 13% to 17% unique variance). It also predicted measures of child speech production: quality (but not quantity) of child speech at 18, 24, and 36 months, quantity (but not quality) of child speech at 24 months (predicting 4% to 14% unique variance). Thus, consistent with prior findings, maternal education plays an increasingly important role in fostering child language development across the first 3 years of life (Hart & Risley, 1995; Hoff, 2003; Rowe, 2018). Critically, accuracy of intersensory matching of faces and voices at 6 months predicted child language outcomes at 18, 24, and 36 months, holding maternal education constant.

Parent Language Input at Older Ages (But Not 6 Months) Predicts Child Language Outcomes

Overall, parent language input (quality and quantity) at 6 months was not a significant predictor of child language outcomes, holding accuracy and speed of intersensory matching and maternal education (SES) constant. The only exception was that quality (but not quantity) of parent language input at 6 months predicted expressive vocabulary at 36 months (predicting 4% unique variance). In contrast, our supplemental analyses revealed that parent language input at older ages (24 and 36 months, but not 18 months) was a significant predictor of child language outcomes, holding other predictors constant. Findings are consistent with previous literature indicating that parent language input at older ages is a strong predictor of child language (Edgar et al., under review; Gilkerson et al., 2018; Hoff & Naigles, 2002; Jones & Rowland, 2017; Pan et al., 2005), and that quality of parent language is a stronger predictor of language outcomes than quantity of parent language at older ages (Hsu et al., 2017; Huttenlocher et al., 1991, 2010; Jones & Rowland, 2017; Rowe, 2012). The current findings also replicate and extend our prior findings that parent language input at older ages predicts child language outcomes (Edgar et al., under review). Thus, given similar levels of intersensory matching skills at 6 months and maternal education, parents who provided greater language input at older ages have children with larger vocabulary size.

Implications for the Study of Language Development

Findings from the present study have a number of implications for the study of child language development. First, the present study adds to a growing body of literature highlighting the importance of assessing individual differences in intersensory processing

for understanding relations between this basic, foundational skill and more complex developmental outcomes. It replicates and extends prior findings demonstrating that intersensory processing of social events (faces and voices) predicts concurrent and future language outcomes in typically developing children (Bahrick, Todd, et al., 2018; Edgar et al., under review) and children with ASD (Todd & Bahrick, under review; (Righi et al., 2018). Second, it highlights the importance of infancy (6 to 12 months) as a foundational period for the development of intersensory processing of social events. Though intersensory processing continues to improve with age, our findings suggest that more efficient selective attention to audiovisual speech in infancy may allow infants to take better advantage of early word learning opportunities (e.g., object labelling), which occur in the context of early social-communicative interactions with caregivers. Third, findings highlight the importance of characterizing developmental pathways and cascades from basic intersensory processing to later, more complex language skills that rely on this foundation. Future research should characterize the potential mediating and moderating roles of parent language input (amount and diversity of child-directed speech) and other variables such as infant speech-like vocalizations, infant bids for attention and engagement in joint attention, and early speech processing efficiency in the context of developmental pathways from intersensory processing of faces and voices to language outcomes. Finally, our findings suggest that impairments of intersensory processing of faces and voices in infancy may be an indicator of risk for language delays. A goal of future research should be to characterize whether early individual differences in intersensory processing of faces and voices can identify children who go on to develop impaired language outcomes. If such links are established, interventions to train and

improve early intersensory processing skills in infancy may be designed and lead to subsequent improvements in later language outcomes.

References

Bahrick, L. E. (1983). Infants' perception of substance and temporal synchrony in multimodal events. *Infant Behavior and Development*, *6*, 429–451. https://doi.org/10.1016/S0163-6383(83)90241-2

Bahrick, L. E. (1987). Infants' intermodal perception of two levels of temporal structure in natural events. *Infant Behavior and Development, 10*(4), 387–416. https://doi.org/10.1016/0163-6383(87)90039-7

Bahrick, L. E. (1988). Intermodal learning in infancy: learning on the basis of two kinds of invariant relations in audible and visible events. *Child Development, 59*(1), 197–209. https://doi.org/10.1111/j.1467-8624.1988.tb03208.x

Bahrick, L. E. (1992). Infants' perceptual differentiation of amodal and modality-specific audio-visual relations. *Journal of Experimental Psychology*, *53*, 180–199.

Bahrick, L. E. (1994). The development of infants' sensitivity to arbitrary intermodal relations. *Ecological Psychology*, *6*(2), 111–123. https://doi.org/10.1207/s15326969eco0602

Bahrick, L. E. (2001). Increasing specificity in perceptual development: Infants' detection of nested levels of multimodal stimulation. *Journal of Experimental Child Psychology*, *79*, 253–270. https://doi.org/10.1006/jecp.2000.2588

Bahrick, L. E. (2010). Intermodal perception and selective attention to intersensory redundancy: Implications for typical social development and autism. In G. Bremner & T. D. Wachs (Eds.), *Blackwell handbook of infant development, 2nd ed.* (Vol. 1, pp. 120–166). Blackwell Publishing. https://doi.org/10.1002/9781444327564.ch4

Bahrick, L. E., Flom, R., & Lickliter, R. (2002). Intersensory redundancy facilitates discrimination of tempo in 3-month-old infants. *Developmental Psychobiology, 41*(4), 352–363. https://doi.org/10.1002/dev.10049

Bahrick, L. E., & Lickliter, R. (2000). Intersensory redundancy guides attentional selectivity and perceptual learning in infancy. *Developmental Psychology, 36*(2), 190–201. https://doi.org/10.1037/0012-1649.36.2.190

Bahrick, L. E., & Lickliter, R. (2002). Intersensory redundancy guides early perceptual and cognitive development. In R. v. Kail (Ed.), *Advances in child development and behavior* (pp. 153–187). Academic Press. https://doi.org/10.1016/S0065-2407(02)80041-6

Bahrick, L. E., & Lickliter, R. (2004). Infants' perception of rhythm and tempo in unimodal and multimodal stimulation: A developmental test of the intersensory redundancy hypothesis. *Cognitive, Affective, and Behavioral Neuroscience*, *4*(2), 137–147.

Bahrick, L. E., & Lickliter, R. (2012). The role of intersensory redundancy in early perceptual, cognitive, and social development. In A. Bremner, D. J. Lewkowicz, & C. Spence (Eds.), *Multisensory development* (pp. 183–205). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199586059.003.0008

Bahrick, L. E., Lickliter, R., & Todd, J. T. (2020). The development of multisensory attention skills: Individual differences, developmental outcomes, and applications. In J. J. Lockman & C. S. Tamis-LeMonda (Eds.), *The Cambridge Handbook of Infant Development* (pp. 303–338). Cambridge University Press.

Bahrick, L. E., Netto, D., & Hernandez-Reif, 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

Bahrick, L. E., & Pickens, J. N. (1988). Classification of bimodal English and Spanish language passages by infants. *Infant and Child Development*, *11*(3), 277–296. https://doi.org/10.1016/0163-6383(88)90014-8

Bahrick, L. E., Soska, K. C., & Todd, J. T. (2018). Assessing individual differences in the speed and accuracy of intersensory processing in young children: The Intersensory Processing Efficiency Protocol. *Developmental Psychology*, *54*(12), 2226–2239. https://doi.org/10.1037/dev0000575

Bahrick, L. E., & Todd, J. T. (2012). Multisensory processing in autism spectrum disorders: Intersensory processing disturbance as a basis for atypical development. In B. E. Stein (Ed.), *The new handbook of multisensory processes* (pp. 657–674). MIT Press.

Bahrick, L. E., Todd, J. T., & Soska, K. C. (2018). The Multisensory Attention Assessment Protocol (MAAP): Characterizing individual differences in multisensory attention skills in infants and children and relations with language

and cognition. *Developmental Psychology*, *54*(12), 2207–2225. https://doi.org/10.1037/dev0000594

Bornstein, M. H., Haynes, M. O., & Painter, K. M. (1998). Sources of child vocabulary competence: A multivariate model. *Journal of Child Language*, *25*(2), 367–393. https://doi.org/10.1017/S0305000998003456

Bremner, A. J. , L. D. J. , & S. C. (2012). *Multisensory development*. Oxford University Press.

Caron, A. J., Caron, R. F., & Maclean, D. J. (1988). Infant discrimination of naturalistic emotional expressions: The role of face and voice. *Child Development*, *59*(3), 604–616.

Dunn, L. M., & Dunn, D. M. (2007). *Peabody Picture Vocabulary Picture (4th ed.)*. NCS Pearson.

Edgar, E. V., Todd, J. T., & Bahrick, L. E. (n.d.). *Intersensory matching of faces and voices in infancy predicts language outcomes in young children*. Manuscript under review.

Fenson, L., Marchman, V. A., Thal, D. J., Dale, P. S., Reznick, J. S., & Bates, E. (2007). *MacArthur-Bates Communicative Development Inventories: User's guide and technical manual (2nd ed.)*. Brookes.

Fernald, A., Perfors, A., & Marchman, V. A. (2006). Picking up speed in understanding: Speech processing efficiency and vocabulary growth across the 2nd year. *Developmental Psychology*, *42*(1), 98–116. https://doi.org/10.1037/0012-1649.42.1.98

Fiebelkorn, I. C., Foxe, J. J., & Molholm, S. (2012). Attention and mulitsensory feature integration. In B. E. Stein (Ed.), *The new handbook of multisensory processing* (pp. 383–394). MIT Press.

Gibson, E. J. (1969). *Principles of Perceptual Learning and Development*. Appleton-Century-Crofts.

Gilkerson, J., Richards, J. A., Warren, S. F., Oller, D. K., Russo, R., & Vohr, B. (2018). Language experience in the second year of life and language outcomes in late childhood. *Pediatrics*, *142*(4). https://doi.org/10.1542/peds.2017-4276

Gogate, L. J., & Bahrick, L. E. (1998). Intersensory redundancy facilitates learning of arbitrary relations between vowel sounds and objects in seven-month-old infants. *Journal of Experimental Child Psychology*, *69*(2), 133–149. https://doi.org/10.1006/jecp.1998.2438

Gogate, L. J., Bahrick, L. E., & Watson, J. D. (2000). A study of multimodal motherese: The role of temporal synchrony between verbal labels and gestures. *Child Development, 71*(4), 878–894. https://doi.org/10.1111/1467-8624.00197

Gogate, L. J., Bolzani, L. H., & Betancourt, E. A. (2006). Attention to maternal multimodal naming by 6- to 8-month-old infants and learning of word-object relations. *Infancy*, *9*(3), 259–288. https://doi.org/10.1207/s15327078in0903_1

Gogate, L. J., & Hollich, G. (2010). Invariance detection within an interactive system: A perceptual gateway to language development. *Psychological Review*, *117*(2), 496–516. https://doi.org/10.1037/a0019049

Gogate, L. J., Walker-Andrews, A. S., & Bahrick, L. E. (2001). The intersensory origins of word comprehension: An ecological-dynamic systems view. *Developmental Science*, *4*(1), 1–18. https://doi.org/10.1111/1467-7687.00143

Guellaï, B., Streri, A., Chopin, A., Rider, D., & Kitamura, C. (2016). Newborns' sensitivity to the visual aspects of infant-directed speech: Evidence from point-line displays of talking faces. *Journal of Experimental Psychology: Human Perception and Performance*, *42*(9), 1275–1281. https://doi.org/10.1037/xhp0000208

Hart, B., & Risley, T. R. (1992). American parenting of language-learning children: Persisting differences in family-child interactions observed in natural home environments. *Developmental Psychology*, *28*(6), 1096–1105. https://doi.org/10.1037/0012-1649.28.6.1096

Hart, B., & Risley, T. R. (1995). *Meaningful Differences in the Everyday Experience of Young American Children*. Paul H. Brookes Publishing.

Hoff, E. (2003). The specificity of environmental influence: Socioeconomic status affects early vocabulary development via maternal speech. *Child Development, 74*(5), 1368–1378. https://doi.org/10.1111/1467-8624.00612

Hoff, E., & Naigles, L. (2002). How children use input to acquire a lexicon. *Child Development, 73*(2), 418–433. https://doi.org/10.1111/1467-8624.00415

Hsu, N., Hadley, P. A., & Rispoli, M. (2017). Diversity matters: Parent input predicts toddler verb production. *Journal of Child Language*, *44*(1), 63–86. https://doi.org/10.1017/S0305000915000690

Huttenlocher, J., Haight, W., Bryk, A., Seltzer, M., & Lyons, T. (1991). Early vocabulary growth: Relation to language input and gender. *Developmental Psychology*, *27*(2), 236–248. https://doi.org/10.1037/0012-1649.27.2.236

Huttenlocher, J., Waterfall, H., Vasilyeva, M., Vevea, J., & Hedges, L. v. (2010). Sources of variability in children's language growth. *Cognitive Psychology*, *61*(4), 343–365. https://doi.org/10.1016/j.cogpsych.2010.08.002

Jackson-Maldonado, D., Thal, D., Marchman, V. A., Newton, T., Fenson, L., & Conboy, B. (2003). *MacArthur Inventorios del Desarollo de Habilidades Communicativas: User's guide and technical manual*. Brookes.

Jesse, A., & Johnson, E. K. (2016). Audiovisual alignment of co-speech gestures to speech supports word learning in 2-year-olds. *Journal of Experimental Child Psychology*, *145*, 1–10. https://doi.org/10.1016/j.jecp.2015.12.002

Jones, G., & Rowland, C. F. (2017). Diversity not quantity in caregiver speech: Using computational modeling to isolate the effects of the quantity and the diversity of the input on vocabulary growth. *Cognitive Psychology*, *98*, 1–21. https://doi.org/10.1016/j.cogpsych.2017.07.002

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

Lewkowicz, D. J. (1992). Infants' response to temporally based intersensory equivalence: The effect of synchronous sounds on visual preferences for moving stimuli. *Infant Behavior and Development*, *15*(3), 297–324. https://doi.org/10.1016/0163-6383(92)80002-C

Lewkowicz, D. J. (2000a). Infants' perception of the audible, visible, and bimodal attributes of multimodal syllables. *Child Development*, *71*(5), 1241–1257. https://doi.org/10.1111/1467-8624.00226

Lewkowicz, D. J. (2000b). The Development of Intersensory Temporal Perception: An Epigenetic Systems/Limitations View. *Psychological Bulletin*, *126*(2), 281–308. https://doi.org/10.1037/0033-2909.126.2.281

Lewkowicz, D. J. (2003). Learning and discrimination of audiovisual events in human infants: The hierarchical relation between intersensory temporal synchrony and rhythmic pattern cues. *Developmental Psychology*, *39*(5), 795–804. https://doi.org/10.1037/0012-1649.39.5.795

Lewkowicz, D. J., Leo, I., & Simion, F. (2010). Intersensory perception at birth: Newborns match nonhuman primate faces and voices. *Infancy*, *15*(1), 46–60. https://doi.org/10.1111/j.1532-7078.2009.00005.x

Lewkowicz, D. J., & Marcovitch, S. (2006). Perception of Audiovisual Rhythm and its Invariance in 4- to 10-Month-Old Infants. *Developmental Psychobiology*, *48*(4), 288–300. https://doi.org/http://dx.doi.org/10.1002/dev.20140

MacWhinney, B. (2000). *The CHILDES Project: Tools for Analyzing Talk (third edition): Transcription format and programs (3rd ed.)*. Lawrence Erlbaum Associates Publishers. https://doi.org/10.1162/coli.2000.26.4.657

Malvern, D., & Richards, B. (2012). Measures of lexical richness. In *The Encyclopedia of Applied Linguistics* (pp. 3622–3627). https://doi.org/10.1002/9781405198431.wbeal0755

Marchman, V. A., & Fernald, A. (2008). Speed of word recognition and vocabulary knowledge in infancy predict cognitive and language outcomes in later childhood. *Developmental Science*, *11*(3), 9–16. https://doi.org/10.1111/j.1467-7687.2008.00671.x.Speed

McCarthy, P. M., & Jarvis, S. (2010). MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods*, *42*(2), 381–392. https://doi.org/10.3758/BRM.42.2.381

Pan, B. A., Rowe, M. L., Singer, J. D., & Snow, C. E. (2005). Maternal correlates of growth in toddler vocabulary production in low-income families. *Child Development*, *76*(4), 763–782. https://doi.org/10.1111/1467-8624.00498-i1

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. (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

Richoz, A. R., Quinn, P. C., De Boisferon, A. H., Berger, C., Loevenbruck, H., Lewkowicz, D. J., Lee, K., Dole, M., Caldara, R., & Pascalis, O. (2017). Audio-visual perception of gender by infants emerges earlier for adult-directed speech. *PLoS ONE*, *12*(1), 1–15. https://doi.org/10.1371/journal.pone.0169325

Righi, G., Tenenbaum, E. J., McCormick, C., Blossom, M., Amso, D., & Sheinkopf, S. J. (2018). Sensitivity to audio-visual synchrony and its relation to language abilities in children with and without ASD. *Autism Research*, *11*(4), 645–653. https://doi.org/10.1002/aur.1918

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

Rowe, M. L. (2008). Child-directed speech: Relation to socioeconomic status, knowledge of child development and child vocabulary skill. *Journal of Child Language*, *35*(1), 185–205. https://doi.org/10.1017/S0305000907008343

Rowe, M. L. (2012). A longitudinal investigation of the role of quantity and quality of child-directed speech vocabulary development. *Child Development*, *83*(5), 1762–1774. https://doi.org/10.1111/j.1467-8624.2012.01805.x

Rowe, M. L. (2018). Understanding socioeconomic differences in parents' speech to children. *Child Development Perspectives*, *12*(2), 122–127. https://doi.org/10.1111/cdep.12271

Rowe, M. L., Pan, B. A., & Ayoub, C. (2005). Predictors of variation in maternal talk to children: A longitudinal study of low-income families. *Parenting*, *5*(3), 259–283. https://doi.org/10.1207/s15327922par0503_3

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*(3), 581–592. https://doi.org/10.1093/biomet/63.3.581

Soderstrom, M., Grauer, E., Dufault, B., & McDivitt, K. (2018). Influences of number of adults and adult: child ratios on the quantity of adult language input across childcare settings. *First Language*, *38*(6), 563–581. https://doi.org/10.1177/0142723718785013

Soken, N. H., & Pick, A. D. (1992). Intermodal perception of happy and angry expressive behaviors by seven-month-old infants. *Child Development*, *63*(4), 787–795. https://doi.org/10.1111/j.1467-8624.1992.tb01661.x

Spelke, E. (1976). Infants' intermodal perception of events. *Cognitive Psychology*, *8*, 553–560. https://doi.org/10.1016/0010-0285(76)90018-9

Stevenson, R. A., Segers, M., Ferber, S., Barense, M. D., & Wallace, M. T. (2014). The impact of multisensory integration deficits on speech perception in children with autism spectrum disorders. *Frontiers in Psychology*, *5*, 1–4. https://doi.org/10.3389/fpsyg.2014.00379

Vaillant-Molina, M., Bahrick, L. E., & Flom, R. (2013). Young infants match facial and vocal emotional expressions of other infants. *Infancy*, *18*, 97–111. https://doi.org/10.1111/infa.12017

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. In *Ecological Psychology* (Vol. 3, Issue 2, pp. 55–75). https://doi.org/10.1207/s15326969eco0302_1

Walker-Andrews, A. S., & Grolnick, W. (1983). Discrimination of vocal expressions by young infants*. *Infant Behavior and Development*, *6*(4), 491–498. https://doi.org/10.1016/S0163-6383(83)90331-4

Weisleder, A., & Fernald, A. (2013). Talking to children matters: Early language experience strengthens processing and builds vocabulary. *Psychological Science*, *24*(11), 2143–2152. https://doi.org/10.1177/0956797613488145

Weizman, Z. O., & Snow, C. E. (2001). Lexical input as related to children's vocabulary acquisition: Effects of sophisticated exposure and support for meaning. *Developmental Psychology, 37*(2), 265–279. https://doi.org/10.1037/0012-1649.37.2.265

Williams, K. T. (2007). *Expressive vocabulary test (2nd ed.)* (NCS Pearson, Ed.).

Tables

Table 1

Demographic information for the sample (N = 103).

Gender	***N***	**Percentage**
Male	51	49.5%
Female	52	50.5%
Ethnicity		
Hispanic	66	64%
Non-Hispanic	34	33%
Did not disclose	3	2.9%
Race		
White/European-American	69	67%
Black/African-American	16	15.5%
Asian/Pacific Islander	2	2%
More than 1 race	9	8.7%
Did not disclose	7	6.8%
Maternal Education		
High school or equivalent	14	13.6%
Some college	16	15.5%
Associate's degree	15	14.6%
Bachelor's degree	26	25.2%
Master's degree or higher	26	25.2%
Did not disclose	6	5.8%
Home Language		
English	63	61.2%
Spanish	30	29.1%
Both English and Spanish	1	1%
Did not disclose	5	4.9%
Age	***M***	***SD***
6-month visit	5.97	.20
18-month visit	18.05	.42
24-month visit	24.19	.37
36-month visit	36.13	.64

Table 2

Protocols, assessments used to index each construct, ages administered, and dependent variables.

Construct	**Protocol/Assessment**	**Ages**	**Dependent Variables**
Infant Intersensory Matching	Intersensory Processing Efficiency Protocol (IPEP)	6 months	Accuracy Speed
Parent Language Input	Parent-Child Interaction (PCI)	6 months	Quantity- Tokens Quality- Types
Child Speech Production	Parent-Child Interaction (PCI)	18, 24, 36 months	Quantity- Tokens Quality- Types
Child Vocabulary Size	Mac-Arthur Bates Communicative Development Inventory (CDI)	18, 24 months	Expressive Vocabulary Receptive Vocabulary
	Expressive Vocabulary Test (EVT)	36 months	Expressive Vocabulary
	Peabody Picture Vocabulary Test (PPVT)	36 months	Receptive Vocabulary

Table 3

Means (M), standard deviations (SD), sample sizes (N), and percentages of missing data for 6-month intersensory matching (both speed and accuracy) for social events, and parent language input (both quantity and quality), as well as 18- 24- and 36-month child language outcomes.

	M	***SD***	***N***	**Missing**
6-Month Intersensory Matching				
Accuracy	.17	.04	90	13.5%
Speed	2.60	0.75	90	13.5%
6-Month Parent Language Input				
Quality (Types)	12.26	5.32	84	19.2%
Quantity (Tokens)	40.56	19.79	84	19.2%
Maternal Education	4.33	1.45	97	6.7%
18-Month Child Language Outcomes				
Child Speech Quality (Types)	.65	.75	76	26.9%
Child Speech Quantity (Tokens)	1.50	1.80	76	26.9%
Receptive Vocabulary (CDI)	231.67	148.90	51	51%
Expressive Vocabulary (CDI)	61.75	77.93	51	51%
24-Month Child Language Outcomes				
Child Speech Quality (Types)	2.70	2.09	70	32.7%
Child Speech Quantity (Tokens)	6.13	5.25	70	32.7%
Expressive Vocabulary (CDI)	275.37	179.99	51	51%
36-Month Child Language Outcomes				
Child Speech Quality (Types)	6.18	3.60	76	26.9%
Child Speech Quantity (Tokens	16.08	12.81	76	26.9%
Receptive Vocabulary (PPVT)	108.85	15.68	71	31.7%
Expressive Vocabulary (EVT)	106.85	15.34	67	35.6%

Table 4

Correlations among predictors (accuracy and speed of intersensory matching of social events, quantity and quality of parent language input, and maternal education) at 6 months and child language outcomes at 18, 24, and 36 months.

	18-Month Child Language Outcomes			
Predictors	Types	Tokens	Receptive	Expressive
6-Month Intersensory Matching				
Accuracy	.37***	.33***	-.01	.33***
Speed	.16	.12	-.11	.10
6-Month Parent Language Input				
Quality (Types)	.12	.02	.04	.07
Quantity (Tokens)	.19*f	.02	.03	.12
Maternal Education	.27**	.21*f	.07	.14

	24-Month Child Language Outcomes		
Predictors	Types	Tokens	Expressive
6-Month Intersensory Matching			
Accuracy	.35***	.40***	.25**
Speed	.10	.07	-.01
6-Month Parent Language Input			
Quality (Types)	.28**	.21*	.22*
Quantity (Tokens)	.24*	.21*	.23*
Maternal Education	.46***	.34***	.26**

	36-Month Child Language Outcomes			
Predictors	Types	Tokens	Receptive	Expressive
6-Month Intersensory Matching				
Accuracy	.15	.05	.37***	.25**
Speed	.02	.03	.03	.09
6-Month Parent Language Input				
Quality (Types)	.15	.09	.22*	.24*
Quantity (Tokens)	.12	.10	.18	.10
Maternal Education	.23*	.09	.40***	.46***

Note: ***p< .001, **p< .01, *p< .05, and *f p < .05 but did not meet significance cut off (p = .0125 for 18 and 36 months, p = .0167 for 24 months) when controlling for familywise error.

Table 5

Amount of unique variance accounted for by each predictor variable (accuracy and speed of intersensory matching for social events, quantity and quality of parent language input, and maternal education) in predicting child language outcomes at 18, 24, and 36 months (N = 103).

	18-Month Language Outcomes			
Predictors	Quantity	Quality	Expressive	Receptive
Total Variance	.18†	.27**	.16	.02
Unique Variance				
6-Month Intersensory Matching				
Accuracy	.11**	.13**	.08*	.00
Speed	.03*	.04*	.04	.02
6-Month Parent Language Input				
Quantity	.00	.02	.00	.00
Quality	.01	.02	.00	.00
Maternal Education	.04	.06*	.01	.01

	24-Month Language Outcomes		
Predictors	Quantity	Quality	Expressive
Total Variance	.29**	.35***	.15
Unique Variance			
6-Month Intersensory Matching			
Accuracy	.14***	.10**	.06*
Speed	.02	.00	.00
6-Month Parent Language Input			
Quantity	.01	.00	.01
Quality	.01	.00	.00
Maternal Education	.08*	.14****	.03

	36-Month Language Outcomes			
Predictors	Quantity	Quality	Expressive	Receptive
Total Variance	.02	.08	.32***	.30**
Unique Variance				
6-Month Intersensory Matching				
Accuracy	.00	.03	.08*	.15***
Speed	.00	.00	.03	.00
6-Month Parent Language Input				
Quantity	.00	.00	.03	.00
Quality	.00	.00	.04*	.00
Maternal Education	.00	.04*	.17**	.13**

Note: ***p< .001, **p< .01, *p< .05

Figures

Figure 0-1. Static image of the dynamic audiovisual social events from the IPEP. On each trial, all six women are shown speaking while the natural and synchronous soundtrack to only one of them is heard. accompanying the videos is synchronized with one of the six women.

Figure 2. Parents and children received three age-appropriate toys during the Parent-Child Interaction (PCI). Each interaction was video recorded by three cameras placed in corners of the playroom (see Edgar et al., under review, for details). Above is a side view of a parent seated across from the 6-month-old infant playing with one of the three toys provided.

I.V. OVERALL CONCLUSION

The manuscripts in this dissertation provide an up-to-date, comprehensive picture of the body of research on intersensory processing of faces and voices in infants and children. In Manuscript 1, we integrated five decades of research that used different measures and paradigms to assess intersensory processing of faces and voices, primarily with a group difference approach. We highlighted a number of convergent findings across paradigms that supported general principles of typical and atypical development. It was evident across several paradigms that temporal synchrony provides a basis for matching faces and voices from a very young age, as well as for detecting nested levels of audiovisual relations that specify properties such as spectral information, affect, prosody, age and gender. Findings converged to demonstrate intersensory processing improvement across age and support for the principle of intersensory facilitation predicted by the IRH. There was clear evidence for both direct and indirect relations between intersensory processing skills and language outcomes. Finally, convergent findings across measures and paradigms demonstrated that infants at risk for developmental delays or children with developmental disabilities show deficits in intersensory processing of faces and voices. This body of research has provided a foundation of knowledge about intersensory processing skills primarily for groups of infants at specific ages. We then discuss the importance of a shift in the field's focus to one of assessing individual differences in intersensory processing skills and their ability to predict later outcomes, including language, social, and cognitive functioning. Two new individual difference measures of intersensory processing have been developed and successfully used to address these questions.

In Manuscript 2, we provided an empirical study illustrating how using one of these individual difference measures can advance the field of intersensory processing. This study was designed to reveal developmental pathways from early intersensory processing skills to later language outcomes. Previously, we found that intersensory processing of faces and voices assessed by the MAAP at 12 months of age predicted unique variance in child language outcomes at 18 and 24 months, over and above the unique variance contributed by traditional predictors, parent language input (quantity and quality) and SES (Edgar et al., under review). Results of the present study extended these findings by using a fine-grained measure of just intersensory processing, the IPEP, with infants of a younger age (6 months), and by predicting language at a later age (36 months). Intersensory processing of faces and voices at 6 months of age predicted unique variance in child language outcomes at 18, 24, and 36 months, over and above the unique variance contributed by parent language input (quantity and quality) and SES. Findings from both empirical studies suggest that 6 to 12 months is an important time in development to assess the role of intersensory processing of faces and voices as a predictor of later child language development.

Future Directions

The use of individual difference measures promises to advance the field of research on intersensory processing, opening the door to addressing a number of important new research questions. First, future research can construct models of developmental growth to identify typical and atypical trajectories of intersensory processing. Given that intersensory processing skills provide a foundation for later, more complex outcomes, it can be used to identify infants at risk for developmental delays. To

accomplish this, a typical trajectory of intersensory processing skills must first be established, and in turn, can then serve as the basis for identifying infants who may be at risk for delays. The infants and children who display intersensory processing skills outside the "typical" range of variability can be identified and targeted for interventions to improve their intersensory functioning.

Second, given the facilitating effect of intersensory redundancy for educating selective attention and promoting learning, future research can assess how to successfully train intersensory processing skills to improve intersensory functioning. To do this, training studies must be designed and assessed for their effectiveness in training intersensory functioning. Further, this area of research would need to assess the generalization of training studies to other social contexts and events. Overall, training intersensory processing skills can foster flexible intersensory functioning and provide a basis for promoting developmental change.

Third, future research can build longitudinal models of developmental pathways to investigate how early developing, intersensory processing skills cascade into later developmental outcomes, such as language. My prior research has demonstrated direct relations between intersensory processing at 6 and 12 months and later language outcomes while controlling for parent language input (quantity and quality) and SES. In contrast, few studies have explicitly demonstrated how intersensory processing transforms and cascades into later developing, more complex skills. Some studies suggest that intersensory processing cascade into skills such as word-mapping, speech processing efficiency, and joint attention, among others, which in turn predict child language (Bahrick & Todd, 2012; Flom & Bahrick, 2007; Gogate & Bahrick, 1998). An ongoing

project is examining the relations among infant intersensory processing skills, quantity and quality of parent language input, basic measures of language learning (e.g., infant speech-like and canonical vocalizations), and later language outcomes to begin to explicitly assess these cascades.

Finally, longitudinal studies can be conducted to investigate how intersensory processing skills change across age in conjunction with skills in other domains (e.g., social and cognitive skills) and aspects of the infant's environment (e.g., parent language input, home language exposure). Longitudinal studies using individual difference measures allow for the examination of the dynamic growth of intersensory processing skills in an embedded environmental context. I am currently conducting two projects using this approach to follow up on my dissertation research. The first project investigates the changing relations among intersensory processing of social events and parent language input across the first two years of life in predicting child language outcomes. The second project assesses the influence of home language exposure on intersensory processing skills of social events across the first three years of life. There are a range of future directions for research using this approach.

Overall, there are a number of fruitful avenues for future research using longitudinal designs and individual difference measures to study intersensory processing. This dissertation advanced our understanding of the body of research on intersensory processing to reveal many of these important new directions for future research. It also demonstrated the powerful role of intersensory processing as a foundation for child language development.

References

Bahrick, L. E., & Todd, J. T. (2012). Multisensory processing in autism spectrum disorders: Intersensory processing disturbance as a basis for atypical development. In B. E. Stein (Ed.), *The new handbook of multisensory processes* (pp. 657–674). MIT Press.

Flom, R., & Bahrick, L. E. (2007). The development of infant discrimination of affect in multimodal and unimodal stimulation: The role of intersensory redundancy. *Developmental Psychology*, *43*(1), 238–252. https://doi.org/10.1037/0012-1649.43.1.238

Gogate, L. J., & Bahrick, L. E. (1998). Intersensory redundancy facilitates learning of arbitrary relations between vowel sounds and objects in seven-month-old infants. *Journal of Experimental Child Psychology*, *69*(2), 133–149. https://doi.org/10.1006/jecp.1998.2438

APPENDIX

MANUSCRIPT TWO: SUPPLEMENTAL MATERIAL

Manuscript Two: Supplemental Material

The present study examined infant intersensory processing of social events (as assessed by the Intersensory Processing Efficiency Protocol; IPEP), SES (as assessed by maternal education) and parent language input (quantity and quality) at 6 months as predictors of child language outcomes (quantity and quality of child speech production, child expressive and receptive vocabulary) at 18, 24, and 36 months.

Method

Child Intersensory Processing Measures: IPEP

Eye-tracking and Data Processing

The IPEP was used to assess accuracy and speed of intersensory matching (see Manuscript, pp. 12-15). Infant eye gaze was sampled at 120Hz by the Tobii X120 system. A Velocity-Threshold Identification (I-VT) filter was used to derive fixations from the raw gaze data (for details see, Olsen, 2012). Six areas of interest (AOIs) were defined within the 2 X 3 grid demarcating each of the six concurrent events on each social and nonsocial trial. The length of each fixation and whether it fell within each AOI or off-AOI was derived from the filtered fixation data and matched to the target and distractor locations based on the target AOI on each trial. For additional details regarding eye-tracking and data processing, see Bahrick, Soska et al. (2018; p. 2231).

Child Vocabulary Measures

Mac-Arthur Bates Communicative Development Inventory (CDI)

The MB-CDI was used to assess expressive and receptive vocabulary (see Manuscript, p. 16). It was completed in English (18 months, n = 46; 24 months, n = 49), or in both English and Spanish (18 months, n = 12; 24 months, n = 14). The inclusion of

children with versus without MB-CDIs in both English and Spanish did not alter findings of the main analyses. Therefore, these participants were included to maximize power. The Words and Gestures form was administered at 18 months. Parents indicated on a checklist (English = 396 items; Spanish = 428 items) which words their child understands (receptive vocabulary) and which words their child understands and says (expressive vocabulary). At 24 months, the Words and Sentences form was administered. Parents indicated (English = 680 items; Spanish = 680 items) which words their child understands and says (expressive vocabulary). For the children whose parents completed MB-CDIs in both English and Spanish, we calculated the total number of words across both forms. Previous literature indicates that vocabulary size is similar across monolingual and bilingual speaking children when the words for the bilingual speaking children are combined from both languages, yielding a total vocabulary size (Pearson et al., 1997; Ramírez-Esparza et al., 2017).

Results

Effects of Home Language, Ethnicity, Race, and Gender

We assessed home language, gender, race, and ethnicity as covariates in the multiple regression analyses in which intersensory matching was a significant predictor of child language outcomes.

Home language

Home language (only English, only Spanish, both English and Spanish) was assessed as a covariate in the multiple regression analyses (see Manuscript, p. 17) to determine if it impacted main findings. When included as a covariate along with the main regression predictors (accuracy of intersensory matching, speed of intersensory matching,

quality of parent language input, quantity of parent language input, and maternal education), home language significantly predicted only 1 of the 11 total child language outcomes: expressive vocabulary size on the MB-CDI at 24 months, $b = 174.09$, $SE = 54.24$, $p = .01$. Critically, even when covarying for home language, 6-month accuracy of intersensory matching of social events was still a significant predictor of 24-month child expressive vocabulary, $b = 10.99$, $SE = 4.80$, $p = .02$. Thus, the inclusion of home language as a covariate did not qualify the main results of our analyses. Further, 6-month accuracy of intersensory matching of social events remained a significant predictor of the same child language outcomes when controlling for home language (ps < .01).

Gender

Gender was assessed as a covariate in the multiple regression analyses (see Manuscript, p. 17) to examine if it impacted the main findings. When included as a covariate along with the main regression predictors (accuracy of intersensory matching, speech of intersensory matching, quality of parent language input, quantity of parent language input, and maternal education), gender was not a significant predictor of any child language outcomes (quantity or quality of child speech production, child vocabulary size) at 18, 24, or 36 months. Thus, the inclusion of gender as a covariate did not qualify the main results of our analyses. Further, 6-month accuracy of intersensory matching of social events remained a significant predictor of the same child language outcomes when controlling for gender (ps < .02).

Race

Race was assessed as a covariate in the multiple regression analyses (see Manuscript, p. 17) to determine if it impacted the main findings. When included as a

covariate along with the main regression predictors (accuracy of intersensory matching, speed of intersensory matching, quality of parent language input, quantity of parent language input, and maternal education), race significantly predicted 2 of the 11 total child language outcomes: expressive vocabulary size on the MB-CDI at 24 months (b = 121.79, SE = 44.69, p = .006) and EVT scores at 36 months (b = 8.40, SE = 3.36, p = .01). However, even when covarying for race, 6-month accuracy of intersensory matching of social events was still a significant predictor of 24-month child expressive vocabulary (b = 9.48, SE = 4.84, p = .05) and 36-month expressive vocabulary (b = .75, SE = .36, p = .04). Therefore, the inclusion of race as a covariate did not qualify the main results of our analyses. Further, 6-month accuracy of intersensory matching of social events remained a significant predictor of the same child language outcomes when controlling for race (ps < .04).

Ethnicity

Finally, we assessed ethnicity (Hispanic, non-Hispanic) as a covariate in the multiple regression analyses (see Manuscript, p. 17) examine if it impacted the main findings. When included as a covariate along with the main regression predictors (accuracy of intersensory matching, speed of intersensory matching, quality of parent language input, quantity of parent language input, and maternal education), ethnicity significantly predicted 1 of the 11 total child language outcomes: expressive vocabulary size on the MB-CDI at 24 months (b = -38.70, SE = 18.85, p = .04). Children of non-Hispanic ethnicity had about 39 fewer words in their expressive vocabulary than children of Hispanic ethnicity at 24 months, controlling for 6-month accuracy and speed of intersensory matching, quantity of parent language input, quality of parent language

input, and maternal education. When covarying for ethnicity, 6-month accuracy of intersensory matching of social events became a marginal predictor of 24-month expressive vocabulary (b = 8.68, SE = 5.08, p = .09). Further, 6-month accuracy of intersensory matching of social events remained a significant predictor of the other child language outcomes when controlling for ethnicity (ps < .01).

Overall, when home language, gender, race, and ethnicity were included as covariates along with the main regression predictors (accuracy and speed of intersensory matching, quality of parent language input, quantity of parent language input, and maternal education), 6-month accuracy of intersensory matching of social events remained a significant predictor of child language outcomes.

Nonsocial Events: Six-Month Intersensory Matching (Speed and Accuracy) as a Predictor of Child Language Outcomes

The present study focused on speed and accuracy of intersensory matching for social events, given that they feature women speaking, an important language learning context for children. Further, nonsocial events were weak or insignificant predictors of language outcomes in our prior studies (Bahrick et al., 2018; Edgar et al., under review). Thus, all analyses in the main manuscript focus on social events. However, in supplemental analyses we also examined whether speed and accuracy of intersensory matching for nonsocial events on the IPEP predicted child language outcomes.

Nonsocial Events: Correlational Analyses

We first assessed correlations between six-month intersensory matching (both speed and accuracy) for nonsocial events in relation to child language outcomes at 18, 24, and 36 months (see Supplemental Table 1). Accuracy of intersensory matching for

nonsocial events was significantly related to one child language outcome: expressive vocabulary size at 18 months, $r = .26$, $p < .01$. Infants who looked longer at the sound-synchronous object at 6 months had a larger expressive vocabulary size at 18 months. It was not correlated with the other 10 child language outcomes. In contrast, accuracy of intersensory matching for social events at 6 months was significantly correlated with 8 out of the 11 child language outcomes at 18, 24, and 36 months (ps < .01).

Speed of intersensory matching for nonsocial events was significantly related to quantity of child speech production ($r = .58$, $p < .001$) and both receptive ($r = .49$, $p < .001$) and expressive vocabulary size ($r = .35$, $p < .001$) at 36 months. Infants who were *slower* to fixate to the sound-synchronous object at 6 months used a greater number of total words (quantity) when speaking to their parent and had greater receptive and expressive vocabularies at 36 months. Unlike for social events, speed of intersensory matching for nonsocial events was related to 3 of the 11 child language outcomes, and slower speed predicted better language outcomes.

Nonsocial Events: Multiple Regression Analyses

We next conducted multiple regressions to assess if accuracy and speed of intersensory matching of nonsocial events at 6 months remained a significant predictor of child language outcomes at 18, 24, and 36 months, even after holdings parent language input (quality and quantity) and SES (maternal education) constant (see Supplemental Table 2). We conducted multiple regression analyses only for the child language outcomes that were significantly correlated with accuracy or speed of intersensory matching of nonsocial events at 6 months (expressive vocabulary at 18 months, quantity of child speech production at 36 months, and receptive and expressive vocabulary at 36

months). However, because few significant relations emerged, they are not discussed in the manuscript results or conclusions.

Accuracy of intersensory matching for nonsocial events at 6 months did not significantly predict child expressive vocabulary at 18 months, after controlling for quantity and quality of parent language input and maternal education (p = .53). On the other hand, speed of intersensory matching at 6 months remained a significant predictor of quantity of child speech production (p = .01) and both receptive (p = .03) and expressive vocabulary (p = .02) size at 36 months (see Supplemental Table 2). However, these effects were positive, indicating that infants who were slower to fixate to the sound-synchronous object at 6 months used a greater amount of words and had greater receptive and expressive vocabularies at 36 months. It appears that slower speed of intersensory matching for nonsocial events at 6 months of age predicts better child language outcomes at 36 months, even after controlling for quantity and quality of parent language input at 6 months and maternal education.

Social Events: Six-Month Intersensory Matching (Speed and Accuracy) as a Predictor of Child Language Outcomes

Social Events: Correlational Analyses

Correlations were conducted between our main predictor variables—accuracy of intersensory matching for social events at 6 months, speed of intersensory matching at 6 months, quality of parent language input at 6 months, quantity of parent language input at 6 months, and maternal education—and our language outcome variables—quality of child speech production, quantity of child speech production, receptive vocabulary, and

expressive vocabulary—at 18, 24, and 36 months of age (there was no measure of receptive vocabulary size at 24 months; see Manuscript, pp. 18-19).

Correlations Between Intersensory Matching for Social Events and Child Language Outcomes. Accuracy of intersensory matching of social events at 6 months predicted 3 out of the 4 language outcomes at 18 months (quality of child speech production, quantity of child speech production, expressive vocabulary size; *r*s > .33, *p*s < .001, but not receptive vocabulary size), all 3 language outcomes at 24 months (quality of child speech, quantity of child speech, expressive vocabulary; *r*s > .25, *p*s < .01), and 2 out of the 4 language outcomes at 36 months (receptive vocabulary, expressive vocabulary; *r*s > .25, *p*s < .01; but not quality or quantity of child speech). In contrast, speed of intersensory matching of social events was not significantly correlated with any of the child language outcomes at 18, 24, or 36 months.

Correlations Between Parent Language Input and Child Language Outcomes. Both measures of parent language input (quality and quantity) at 6 months significantly predicted all 3 language outcomes at 24 months (quality of child speech, quantity of child speech, expressive vocabulary; *r*s > .21, *p*s < .05), whereas only quality of parent language input (but not quantity) predicted 2 out of the 4 child language outcomes at 36 months (receptive vocabulary, expressive vocabulary; *r*s > .25, *p*s < .01). In contrast, neither quality nor quantity of parent language input was significantly correlated with any child language outcome at 18 months.

Correlations Between Maternal Education and Child Language Outcomes. Maternal education significantly predicted 2 out of the 4 child language outcomes at 18 months (quality of child speech, quantity of child speech; *r*s > .21, *p* < .05), all 3

language outcomes at 24 months (quality of child speech, quantity of child speech, expressive vocabulary; *r*s > .26, *p*s < .01), and 3 out of the 4 language outcomes at 36 months (quality of child speech, receptive vocabulary, expressive vocabulary; *r*s > .23, *p*s < .05; but not quantity of child speech).

Social Events: Multiple Regression Analyses

For each of the 11 outcome variables, we conducted five multiple regression models to assess the amount of unique variance (ΔR^2) attributable to each predictor in predicting the outcome variable (see Manuscript, pp. 19-22). The unique variance attributable to a given predictor is the change in R^2 when that predictor is entered last in the regression model (i.e., holding all other predictors constant). The regression coefficients and unique variance attributable to each predictor can be found in Supplemental Tables 4 through 9. The amount of total variance explained by all 5 predictors, as well as the unique variance explained by each predictor in predicting each of the language outcomes at each age are detailed below and are summarized in the Manuscript, Table 5.

18-Month Child Speech Production (Quality, Quantity). Together, all 6-month predictors accounted for a significant 27% of the total variance in child speech quality (types), $p < .001$, and a marginal 18% of the total variance in child speech quantity (tokens), $p < .10$, at 18 months (see Supplemental Table 4). Accuracy of intersensory matching of social events at 6 months accounted for a significant 13% and 11% unique variance in child speech quality and quantity respectively, *p*s < .01. Speed of intersensory matching for social events at 6 months accounted for a smaller but still significant 4% and 3% unique variance in child speech quality and quantity, respectively,

*p*s < .05. Maternal education accounted for a significant 6% unique variance in child speech quantity, $p < .05$ (quality: 4%, *n.s.*), whereas parent language input accounted for non-significant unique variance in child speech (quality: 0%; quantity: 2%, *n.s.*).

18-Month Expressive and Receptive Vocabulary Size (MB-CDI). Together, all 6-month predictors accounted for a non-significant amount of total variance in expressive and receptive vocabulary at 18 months (16% and 2%, respectively; *n.s.*; see Supplemental Table 5). However, accuracy of intersensory matching at 6 months accounted for a significant 8% unique variance in expressive vocabulary, $p < .05$ (receptive: 0%, *n.s.*). All other predictors (speed of intersensory matching, parent language quality and quantity, maternal education) accounted for non-significant amounts of unique variance in vocabulary size (range: 0% to 4%, *n.s.*).

24-Month Child Speech Production (Quality, Quantity). Together, all 6-month predictors accounted for a significant 35% of the total variance in child speech quality (types), $p < .01$, and a significant 29% of total variance in child speech quantity (tokens), $p < .01$, at 24 months (see Supplemental Table 6). Accuracy of intersensory matching of social events at 6 months accounted for a significant 10% and 14% of unique variance in child speech quality and quantity, respectively, *p*s < .01. Also, maternal education accounted for a significant 14% and 8% unique variance in child speech quality and quantity, respectively, *p*s < .05. All other predictors (speed of intersensory matching, parent language quality and quantity) accounted for non-significant amounts of unique variance in child speech production (range: 0% to 2%, *n.s.*).

24-Month Expressive Vocabulary Size (MB-CDI). Together, all 6-month predictors accounted for a non-significant amount of total variance in expressive

vocabulary at 24 months (15%, *n.s.*; see Supplemental Table 7). However, accuracy of intersensory matching of social events at 6 months accounted for a significant 6% unique variance in expressive vocabulary, $p < .05$. All other predictors (speed of intersensory matching, parent language quality and quantity, maternal education) accounted for non-significant amounts of unique variance in expressive vocabulary (range: 0% to 3%, *n.s.*).

36-Month Child Speech Production (Quality, Quantity). Together, all 6-month predictors accounted for non-significant amounts of total variance in child speech quality and quantity at 36 months (8% and 2%, respectively, *n.s.*; see Supplemental Table 8). However, maternal education accounted for a significant 4% unique variance in child speech quality, $p < .05$ (quantity: 0%, *n.s.*). All other predictors (accuracy and speed of intersensory matching, parent language quality and quantity) accounted for non-significant amounts of unique variance in child speech quality and quantity (0%, *n.s.*).

36-Month Expressive and Receptive Vocabulary Size (EVT, PPVT). Together, all 6-month predictors accounted for a significant 32% of the total variance in expressive vocabulary, $p < .001$, and a significant 30% of the total variance in receptive vocabulary, at 36 months, $p < .01$ (see Table 9). Accuracy of intersensory matching of social events at 6 months accounted for a significant 8% and 15% unique variance in child expressive and receptive vocabulary, respectively, $ps < .05$. Maternal education also accounted for a significant 17% and 13% of unique variance in expressive and receptive vocabulary, respectively, $ps < .01$. Further, parent language quantity (but not quality, *n.s.*) accounted for a significant 4% unique variance in expressive vocabulary, $p < .05$ (receptive: 0%, *n.s.*).

Social Events: Quantifying Relations Between 6-Month Accuracy and Speed of Intersensory Matching and Child Language Outcomes

We also inspected the unstandardized coefficients from the multiple regression models presented in Supplemental Tables 4 to 9 to quantify the magnitude of the relationship between accuracy and speed of intersensory matching of social events and later child language outcomes. The unstandardized regression coefficient is an estimate of the average or expected raw score change in an outcome variable (child language outcomes) for each raw score unit increase in a predictor variable (intersensory matching), holding all other predictors constant (parent language input, maternal education; Cohen et al., 2003).

Here we report average change in child language outcomes associated with a 5% increase in accuracy of intersensory matching (roughly equal to the *SD* of accuracy: *SD* = 4%) and change associated with a 1-s decrease in speed of intersensory matching (slightly larger than the *SD* of .75 s; see Manuscript, Table 3). For the 18-month child language outcomes, on average, holding other predictors constant, a 5% increase in in accuracy of intersensory matching was associated with a significant 0.70-word per-minute increase in child speech quantity (tokens), a significant 0.30-word per-minute increase in child speech quality (types), and a marginal 30.45-word per-minute increase in expressive vocabulary (see Supplemental Tables 4 and 5). Holding other predictors constant, a 1-s increase in speed of intersensory matching of social events was associated with a significant 0.44-word per-minute increase in child speech quantity, and a significant 0.20-word per-minute increase in child speech quality. For the 24-month child language outcomes, on average, holding other predictors constant, a 5% increase in

accuracy of intersensory matching was associated with a significant 2.35-word per-minute increase in child speech quantity (tokens), a significant 0.85-word per-minute increase in child speech quality (types), and a significant 48.55-word increase in expressive vocabulary (see Supplemental Tables 6 and 7). For the 36-month child language outcomes, on average, holding other predictors constant, a 5% increase in accuracy of intersensory matching was associated with a 6.65-standard score increase in receptive vocabulary (PPVT), and a 4.30-standard score increase in expressive vocabulary (EVT; see Supplemental Tables 8 and 9). Speed of intersensory matching was not a significant predictor of 24- or 36-month child language outcomes. Overall, improvements in accuracy of intersensory matching of social events at 6 months of age accounted for significant and meaningful improvements in many of the child language outcomes at 18, 24, and 36 months.

Secondary Analyses: Parent Language Input at Older Ages

Findings from our primary analyses demonstrate that intersensory matching (both speed and accuracy) of social events at 6 months was a significant predictor of child language outcomes at 18, 24, and 36 months, after controlling for parent language input (both quantity and quality) at 6 months and maternal education (see Manuscript, pp. 19-22 and Table 10; see Supplemental Tables 4 through 9). However, we also assessed whether speed and accuracy of intersensory matching for social events at 6 months would still predict child language outcomes, controlling for parent language input when children were older (18, 24, and 36 months; See Manuscript, p. 22). In our prior study (Edgar et al., under review) parent language input at older ages (18 and 24 months) was a better predictor of child language outcomes than parent input at 12 months. Other research also

indicates that parent language input at older ages predicts child language outcomes (Gilkerson et al., 2018; Jones & Rowland, 2017; Pan et al., 2005).

Correlations between parent language input (both quantity and quality) at each age and child language outcomes can be found in Supplemental Table 3. Overall, parent language input (quantity and quality) at each age (18, 24, and 36 months) predicted concurrent child language outcomes at 18, 24, and 36 months. Results of the multiple regression analyses for each outcome variable separately are presented in Supplemental Tables 10 through 16.

Secondary Analyses: Intersensory Matching for Social Events at 6 Months Remains a Significant Predictor of Child Language Outcomes When Controlling for Parent Language Input at Older Ages (18, 24, and 36 Months)

After controlling for quantity and quality of parent language input at 18 months, accuracy of intersensory matching for social events at 6 months predicted three of the four language outcomes (quantity and quality of child speech and expressive vocabulary, but not receptive vocabulary) at 18 months, and speed of intersensory matching predicted two (child quantity and quality of speech; see Supplemental Tables 10 and 11). Accuracy of intersensory matching for social events at 6 months accounted for 13% to 17% of the unique variance in child language outcomes, whereas speed of intersensory matching accounted for 2% to 4%. Controlling for parent language input (both quantity and quality) at 24 months, accuracy of intersensory matching for social events at 6 months predicted all three child language outcomes (quantity and quality of child speech productive and expressive vocabulary) at 24 months, while speed of intersensory matching did not significantly predict any of the three outcomes (see Supplemental

Tables 12 and 13). Six-month accuracy of intersensory matching accounted for 7% to 16% of the unique variance in child language outcomes at 24 months. Finally, controlling for parent (both quantity and quality) at 36 months, accuracy and speed of intersensory matching for social events at 6 months predicted two out of four child language outcomes (expressive and receptive vocabulary) at 36 months, while speed of intersensory matching did not significantly predict any of the four outcomes (see Supplemental Tables 14 and 15). Accuracy of intersensory matching accounted for 5% to 12% of the unique variance in child vocabulary at 36 months. In sum, 6-month intersensory processing still predicts language outcomes at 18, 24, and 36 months over and above SES and the language input children receive at those ages.

Secondary Analyses: Parent Language Input at Older Ages (18, 24, and 36 months) Predicts Child Language Outcomes

Multiple regression analyses indicated that quantity and quality of parent language input at 18, 24, and 36 months predict some child language outcomes, after controlling for speed and accuracy of intersensory matching for social events at 6 months and maternal education. After controlling for speed and accuracy of intersensory matching for social events at 6 months and maternal education, quantity, but not quality, of parent language input at 24 months only predicted child expressive vocabulary at 24 months, accounting for 16% of the unique variance. Quality, but not quantity, of parent language input at 36 months significantly predicted child expressive and receptive vocabulary at 36 months, after controlling for speed and accuracy of intersensory matching for social events at 6 months and maternal education. It accounted for 3% to 6% of the unique variance in child expressive and receptive vocabulary at 36 months.

Quantity and quality of parent language input at 18 months did not predict child language outcomes (child quality and quantity of speech, expressive and receptive vocabulary) at 18 months, after controlling for speed and accuracy of intersensory matching at 6 months and maternal education. Thus, quantity of parent language input at 24 months (but not 18 or 36 months) and quality of parent language input at 36 months (but not 18 or 24 months) appear to be strong predictors of some language outcomes (expressive and receptive vocabulary), and a weaker predictor of other outcomes (quantity and quality of child speech), after holding constant speed and accuracy of intersensory matching at 6 months and maternal education.

References

Bahrick, L. E., Todd, J. T., & Soska, K. C. (2018). The Multisensory Attention Assessment Protocol (MAAP): Characterizing individual differences in multisensory attention skills in infants and children and relations with language and cognition. *Developmental Psychology, 54*(12), 2207–2225. https://doi.org/10.1037/dev0000594

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analyses for the behavioral sciences (3rd ed.)*. Lawrence Erlbaum Associates.

Edgar, E. V., Todd, J. T., & Bahrick, L. E. (n.d.). *Intersensory matching of faces and voices in infancy predicts language outcomes in young children*. Manuscript under review.

Gilkerson, J., Richards, J. A., Warren, S. F., Oller, D. K., Russo, R., & Vohr, B. (2018). Language experience in the second year of life and language outcomes in late childhood. *Pediatrics, 142*(4). https://doi.org/10.1542/peds.2017-4276

Jones, G., & Rowland, C. F. (2017). Diversity not quantity in caregiver speech: Using computational modeling to isolate the effects of the quantity and the diversity of the input on vocabulary growth. *Cognitive Psychology, 98*, 1–21. https://doi.org/10.1016/j.cogpsych.2017.07.002

Olsen, A. (2012). The Tobii I-VT Fixation Filter: Algorithm description. In *Tobii Technology*.

Pan, B. A., Rowe, M. L., Singer, J. D., & Snow, C. E. (2005). Maternal correlates of growth in toddler vocabulary production in low-income families. *Child Development, 76*(4), 763–782. https://doi.org/10.1111/1467-8624.00498-i1

Pearson, B. Z., Fernandez, S. C., Lewedeg, V., & Oller, D. K. (1997). The relation of input factors to lexical learning by bilingual infants. *Applied Psycholinguistics, 18*(1), 41–58. https://doi.org/10.1017/s0142716400009863

Ramírez-Esparza, N., García-Sierra, A., & Kuhl, P. K. (2017). The impact of early social interactions on later language development in Spanish–English bilingual infants. *Child Development, 88*(4), 1216–1234. https://doi.org/10.1111/cdev.12648

Supplemental Table 1

Correlations between accuracy and speed of intersensory matching for nonsocial events at 6 months and child language outcomes at 18, 24, and 36 months (N =103).

	18-Month Child Language Outcomes			
6-Month	Quality	Quantity	Receptive	Expressive
Intersensory Matching				
Accuracy	.06	-.19	-.08	.26**
Speed	-.20*f	.001	-.04	.23*f
	24-Month Child Language Outcomes			
	Quality	Quantity	Expressive	
Intersensory Matching				
Accuracy	-.07	-.12	.10	
Speed	.07	.17	.01	
	36-Month Child Language Outcomes			
	Quality	Quantity	Receptive	Expressive
Intersensory Matching				
Accuracy	-.002	-.09	-.08	-.10
Speed	-.05	.58***	.49***	.35***

Note: *p<.05, **p< .01, ***p< .001, and *f: p < .05 but did not meet significance cut off (p = .0125 for 18 and 36 months, p = .0167 for 24 months) when controlling for familywise error.

Supplemental Table 2

Multiple regressions for accuracy and speed of intersensory matching for nonsocial events, quantity and quality of parent language input, and maternal education at 6 months predicting child language outcomes from the significant correlations in Supplemental Table 1. Unstandardized regression coefficients are listed, followed by standard errors in parentheses (N = 103).

	Child Language Outcomes							
	18-Month		**36-Month**					
Predictors	Expressive		Quantity		Receptive		Expressive	
6-Month Intersensory Matching								
Accuracy	.26	(.41)	-.01	(.01)	-.03	(.02)	-.03	(.03)
Speed	1.42	(4.68)	.14**	(.05)	.48*	(.22)	.43*	(.19)
6-Month Parent Language Input								
Quantity (Tokens)	1.12	(4.51)	-.04	(.09)	-.13	(.33)	.01	(.36)
Quality (Types)	35.07	(163.01)	-3.94	(2.94)	-12.67	(10.65)	-3.25	(18.85)
Maternal Education	2.94	(16.55)	.70†	(.39)	1.14	(1.37)	4.31*	(1.67)

Note: *p<.05, **p< .01, ***p< .001

Supplemental Table 3

Correlations between quantity and quality of parent language input at 6, 18, 24, and 36 months and child language outcomes at 18, 24, and 36 months (N = 103).

	Child Language Outcomes										
	18 Months				24 Months			36 Months			
Parent Language	Quality	Quantity	Receptive	Expressive	Quality	Quantity	Expressive	Quality	Quantity	Receptive	Expressive
6 Months											
Quality (Types)	.15	.07	.01	.07	.19	.06	-.03	.19	.13	.18	.19
Quantity (Tokens)	.16	.02	-.04	.02	.13	.05	.08	.12	.08	.17	.08
18 Months											
Quality (Types)	.31**	.28**	.22^{*f}	.25*	.27*	.08	.11	.12	.03	.08	.15
Quantity (Tokens)	.23^{*f}	.23^{*f}	.16	.25*	.18	.06	.23^{*f}	.05	-.003	.12	.01
24 Months											
Quality (Types)	.24^{*f}	.13	.35***	.44***	.41***	.29**	.25*	.23^{*f}	.17	.32***	.31**
Quantity (Tokens)	.22^{*f}	.12	.23^{*f}	.39***	.31**	.23^{*f}	.44***	.14	.09	.32***	.24^{*f}
36 Months											
Quality (Types)	.35***	.31**	.25*	.36***	.37**	.21^{*f}	.17	.33***	.19	.39***	.36***
Quantity (Tokens)	.21^{*f}	.15	.13	.24^{*f}	.17	.05	.24*	.28**	.22^{*f}	.28**	.17

Note: $*p<.05$, $**p< .01$, $***p< .001$, and $*f$ $p < .05$ but did not meet significance cut off ($p = .0125$ for 18 and 36 months, $p = .0167$ for 24 months) when controlling for familywise error.

Supplemental Table 4

Multiple regressions and change in R^2 for accuracy and speed of intersensory matching for social events, parent language input (both quantity and quality), and maternal education at 6 months predicting child speech production at 18 months (N = 103). The unique variance (ΔR^2) for each predictor (when holding all other predictors constant) is presented in Step 5 of each model.

	18-Month Child Speech Quantity							**18-Month Child Speech Quality**						
	Variance		**beta**					**Variance**		**beta**				
Steps and predictors: **6-Month**	Total R^2	ΔR^2	Step 1	Step 2	Step 3	Step 4	Step 5	Total R^2	ΔR^2	Step 1	Step 2	Step 3	Step 4	Step 5
Model 1														
1. Maternal Education	.05	.05	.27	.30	.31	.32†	.29	.08*	.08	.14*	.14*	.15*	.16*	.14*
2. Parent Language: Quality	.05	0	-	-.02	-.04	-.04	-.03	.08	0	-	.004	-.04	-.04	-.04
3. Parent Language: Quantity	.05	0	-	-	.01	.002	-.001	.12	.04	-	-	.01	.01	.01
4. Speed	.07	.02	-	-	-	.35	.44*	.14	.02	-	-	-	.16	.20*
5. Accuracy	**.18**†	**.11****	-	-	-	-	.14**	**.27****	**.13*****	-	-	-	-	.06***
Model 2														
1. Parent Language: Quality	0	0	-.01	-.01	-.002	-.003	-.03	.01	.01	.01	-.03	-.02	-.02	-.04
2. Parent Language: Quantity	0	0	-	.002	-.002	-.004	-.001	.04	.03	-	.01	.01	.01	.01
3. Speed	.02	.02	-	-	.33	.42†	.44*	.06	.02	-	-	.15	.19*	.20*
4. Accuracy	.14†	.12**	-	-	-	.15**	.14**	.21*	.15***	-	-	-	.07***	.06***
5. Maternal Education	**.18**†	**.04**	-	-	-	-	.29	**.27****	**.06***	-	-	-	-	.14*
Model 3														
1. Parent Language: Quantity	0	0	-.001	-.002	-.01	-.01	-.001	.03	.03	.01	.01	.01	.003	.01
2. Speed	.02	.02	-	.32	.41†	.44*	.44*	.05	.02	-	.15	.20*	.21*	.20*
3. Accuracy	.14†	.12**	-	-	.15**	.14**	.14**	.20**	.15***	-	-	.07***	.07***	.06***
5. Maternal Education	.17†	.03	-	-	-	.27	.29	.25**	.05*	-	-	-	.12†	.14*
5. Parent Language: Quality	**.18**†	**.01**	-	-	-	-	-.03	**.27****	**.02**	-	-	-	-	-.04
Model 4														
1. Speed	.02	.02	.31	.40†	.42†	.43†	.44*	.03	.03	.17	.21†	.22*	.22*	.20*
2. Accuracy	.14†	.12**	-	.14***	.14***	.14**	.14**	.19**	.16***	-	.07***	.07***	.07***	.06***
3. Maternal Education	.17†	.03	-	-	.24	.29	.29	.25**	.06*	-	-	.13*	.13*	.14*
4. Parent Language: Quality	.18†	.01	-	-	-	-.03	-.03	.25**	0	-	-	-	.001	-.04

5. Parent Language: Quantity	**.18**†	**0**	-	-	-	-	-.001	**.27****	**.02**	-	-	-	-	.01
Model 5														
1. Accuracy	.11**	0	.14**	.13**	.13**	.13**	.14**	.14*	.14**	.07***	.06***	.06***	.06***	.06***
2. Maternal Education	.14	.03	-	.23	.27	.27	.29	.20*	.06*	-	.13*	.12†	.13*	.14*
3. Parent Language: Quality	.15	.01	-	-	-.02	-.04	-.03	.20*	0	-	-	.004	-.04	-.04
4. Parent Language: Quantity	.15	0	-	-	-	.004	-.001	.23**	.03	-	-	-	.01	.01
5. Speed	**.18**†	**.03***	-	-	-	-	.44*	**.27****	**.04***	-	-	-	-	.20*

Note: $^{\dagger}p<.10$, $^{*}p<.05$, $^{**}p< .01$, $^{***}p< .001$

Supplemental Table 5

Multiple regressions and change in R^2 for accuracy and speed of intersensory matching for social events, parent language input (both quantity and quality), and maternal education at 6 months predicting child vocabulary size at 18 months (N = 103). The unique variance (ΔR^2) for each predictor (when holding all other predictors constant) is presented in Step 5 of each model.

	18-Month Expressive Vocabulary							**18-Month Receptive Vocabulary**						
	Variance		**beta**					**Variance**		**beta**				
Steps and predictors:	Total	ΔR^2	Step	Step	Step	Step	Step	Total	ΔR^2	Step	Step	Step	Step	Step
6-Month	R^2		1	2	3	4	5	R^2		1	2	3	4	5
Model 1														
1. Maternal Education	.03	.03	9.48	9.60	10.79	10.59	7.87	.003	.003	5.31	3.43	3.56	4.79	5.22
2. Parent Language: Quality	.03	0	-	-.21	-5.37	-5.25	-2.35	.003	0	-	1.01	.95	.40	-.02
3. Parent Language: Quantity	.07	.04	-	-	1.52	1.51	.73	.003	0	-	-	.02	.14	.25
4. Speed	.08	.01	-	-	-	6.00	15.25	.02	.017	-	-	-	-20.96	-22.53
5. Accuracy	**.16**	**.08***	-	-	-	-	6.09†	**.02**	**0**	-	-	-		-1.10
Model 2														
1. Parent Language: Quality	.003	.003	.84	-3.83	-3.76	-1.27	-2.35	.001	.001	1.06	.96	.62	.25	-.02
2. Parent Language: Quantity	.04	.037	-	1.40	1.38	.62	.73	.001	0	-	.03	.14	.23	.25
3. Speed	.05	.01	-	-	6.77	16.49	15.25	.01	.009	-	-	-20.56	-21.79	-22.53
4. Accuracy	.15	.10*	-	-	-	6.49*	6.09†	.01	0	-	-	-	-.88	-1.10
5. Maternal Education	**.16**	**.01**	-	-	-	-	7.87	**.02**	**.01**	-	-	-	-	5.22
Model 3														
1. Parent Language: Quantity	.02	.02	.54	.53	.32	.21	.73	.001	.001	.25	.28	.28	.24	.25
2. Speed	.03	.01	-	8.32	17.29	16.57	15.25	.01	.009	-	-20.08	-21.28	-22.15	-22.53
3. Accuracy	.16	.13*	-	-	6.65*	6.48*	6.09†	.01	0	-	-	-.97	-1.11	-1.10
5. Maternal Education	.16	0	-	-	-	6.51	7.87	.02	.01	-	-	-	5.18	5.22
5. Parent Language: Quality	**.16**	**0**	-	-	-	-	-2.35	**.02**	**0**	-	-	-	-	-.02
Model 4														
1. Speed	.01	.01	10.01	18.72	17.63	17.46	15.25	.01	.01	-18.14	-18.38	-19.67	-20.28	-22.53
2. Accuracy	.16	.15*	-	6.93*	6.60*	6.61*	6.09†	.01	0	-	-.56	-.86	-.86	-1.10
3. Maternal Education	.16	0	-	-	7.57	.14	7.87	.02	.01	-	-	6.32	4.90	5.22
4. Parent Language: Quality	.16	0	-	-	-	7.14	-2.35	.02	0	-	-	-	.80	-.02

5. Parent Language: Quantity	**.16**	**0**	-	-	-	-	.73	**.02**	**0**	-	-	-	-	.25
Model 5														
1. Accuracy	.11*	.11*	5.85*	5.57*	5.58*	5.05†	6.09†	0	0	.48	.27	.32	.34	-1.10
2. Maternal Education	.12	.01	-	8.12	7.83	8.72	7.87	.003	.003	-	5.14	3.27	3.39	5.22
3. Parent Language: Quality	.12	0	-	-	.05	-3.15	-2.35	.003	0	-	-	1.01	1.02	-.02
4. Parent Language: Quantity	.12	0	-	-	-	.94	.73	.003	0	-	-	-	-.01	.25
5. Speed	**.16**	**.04**	-	-	-	-	15.25	**.02**	**.017**	-	-	-	-	-22.53

Note: †p<.10, *p<.05, **p< .01, ***p< .001

Supplemental Table 6

Multiple regressions and change in R^2 for accuracy and speed of intersensory matching for social events, parent language input (both quantity and quality), and maternal education at 6 months predicting child speech production at 24 months (N = 103). The unique variance (ΔR^2) for each predictor (when holding all other predictors constant) is presented in Step 5 of each model.

	24-Month Child Speech Quantity							**24-Month Child Speech Quality**						
	Variance		**beta**					**Variance**		**beta**				
Steps and predictors: **6-Month**	Total R^2	ΔR^2	Step 1	Step 2	Step 3	Step 4	Step 5	Total R^2	ΔR^2	Step 1	Step 2	Step 3	Step 4	Step 5
Model 1														
1. Maternal Education	.11**	.11**	1.18*	1.01†	1.10†	1.10†	1.09*	.21***	.21***	.67***	.58**	.60**	.60**	.60**
2. Parent Language: Quality	.11	0	-	.11	-.09	-.09	-.06	.22*	.01	-	.06	.01	.01	.03
3. Parent Language: Quantity	.14	.03	-	-	.06	.07	.04	.23*	.01	-	-	.02	.01	.01
4. Speed	.15	.01	-	-	-	.64	.87	.25*	.02	-	-	-	.38	.46
5. Accuracy	**.29****	**.14*****	-	-	-	-	.47***	**.35*****	**.10****	-	-	-	-	.17***
Model 2														
1. Parent Language: Quality	.06	.06	.23*	.13	.12	.15	-.06	.09*	.09*	.12**	.11	.11	.12	.03
2. Parent Language: Quantity	.06	0	-	.04	.04	.01	.04	.09	0	-	.002	.001	-.01	.01
3. Speed	.08	.02	-	-	.64	.85	.87	.11	.02	-	-	.37	.44	.46
4. Accuracy	.21*	.13***	-	-	-	.47***	.47***	.21*	.10***	-	-	-	.16**	.17***
5. Maternal Education	**.29****	**.08***	-	-	-	-	1.09*	**.35*****	**.14*****	-	-	-	-	.60**
Model 3														
1. Parent Language: Quantity	.06	.06	.07†	.07†	.05	.03	.04	.07	.07	.03†	.03†	.02	.01	.01
2. Speed	.08	.02	-	.58	.79	.87	.87	.09	.02	-	.33	.40	.45	.46
3. Accuracy	.20*	.12***	-	-	.47***	.47***	.47***	.18*	.09**	-	-	.16**	.16***	.17***
5. Maternal Education	.28**	.08*	-	-	-	1.05*	1.09*	.35***	.17***	-	-	-	.62***	.60**
5. Parent Language: Quality	**.29****	**.01**	-	-	-	-	-.06	**.35*****	**0**	-	-	-	-	.03
Model 4														
1. Speed	.01	.01	.72	.90	.94	.90	.87	.02	.02	.37	.44	.46	.45	.46
2. Accuracy	.18*	.17***	-	.50***	.49***	.49***	.47***	.14†	.12**	-	.17**	.17***	.17***	.17***
3. Maternal Education	.28**	.10**	-	-	1.15**	1.04*	1.09*	.35***	.21***	-	-	.66***	.60***	.60**
4. Parent Language: Quality	.28**	0	-	-	-	.07	-.06	.35***	0	-	-	-	.04	.03

5. Parent Language: Quantity	**.29****	**.01**	-	-	-	-	.04	**.35*****	**0**	-	-	-	-	.01
Model 5														
1. Accuracy	.17*	.17*	.49***	.48***	.47***	.46***	.47***	.13	.13	.17**	.17***	.16***	.16**	.17***
2. Maternal Education	.27**	.10**	-	1.14**	1.04*	1.09*	1.09*	.33***	.20***	-	.65***	.59***	.60***	.60**
3. Parent Language: Quality	.27**	0	-	-	.06	-.06	-.06	.33***	0	-	-	.04	.02	.03
4. Parent Language: Quantity	.27**	0	-	-	-	.04	.04	.33***	0	-	-	-	.01	.01
5. Speed	**.29****	**.02**	-	-	-	-	.87	**.35*****	**.02**†	-	-	-	-	.46

Note: $^{\dagger}p<.10$, $^{*}p<.05$, $^{**}p< .01$, $^{***}p< .001$

Supplemental Table 7

Multiple regressions and change in R^2 for accuracy and speed of intersensory matching for social events, parent language input (both quantity and quality), and maternal education at 6 months predicting child vocabulary size at 24 months (N = 103). The unique variance (ΔR^2) for each predictor (when holding all other predictors constant) is presented in Step 5 of each model.

	24-Month Expressive Vocabulary						
	Variance		beta				
Steps and predictors: 6-Month	Total R^2	ΔR^2	Step 1	Step 2	Step 3	Step 4	Step 5
Model 1							
1. Maternal Education	.06	.06	30.59†	22.89	26.17	25.85	25.45
2. Parent Language: Quality	.07	.01	-	4.38	-1.40	-1.69	-.67
3. Parent Language: Quantity	.09	.02	-	-	1.83	1.99	1.66
4. Speed	.09	0	-	-	-	-3.56	2.39
5. Accuracy	**.15**	**.06***	-	-	-	-	9.71*
Model 2							
1. Parent Language: Quality	.05	.05	7.83†	5.19	4.66	5.30	-.67
2. Parent Language: Quantity	.06	.01	-	.89	1.10	.83	1.66
3. Speed	.06	0	-	-	-4.16	1.65	2.39
4. Accuracy	.12	.06*	-	-	-	9.50*	9.71*
5. Maternal Education	**.15**	**.03**	-	-	-	-	25.45
Model 3							
1. Parent Language: Quantity	.06	.06	2.16	2.24	2.13	1.51	1.66
2. Speed	.06	0	-	-5.44	.08	2.68	2.39
3. Accuracy	.11	.05*	-	-	9.37*	9.73*	9.71*
5. Maternal Education	.15	.04	-	-	-	24.89	25.45
5. Parent Language: Quality	**.15**	**0**	-	-	-	-	-.67
Model 4							
1. Speed	.001	.001	-6.43	-.57	2.83	4.13	2.39
2. Accuracy	.06	.059*	-	9.86*	10.0*	9.94*	9.71*
3. Maternal Education	.13	.07*	-	-	30.72†	22.57	25.45
4. Parent Language: Quality	.14	.01	-	-	-	4.58	-.67
5. Parent Language: Quantity	**.15**	**.01**	-	-	-	-	1.66
Model 5							
1. Accuracy	.06	.06	9.87*	9.96*	9.89*	9.71*	9.71*
2. Maternal Education	.13	.07*	-	30.65†	23.07	25.75	25.45
3. Parent Language: Quality	.13	0	-	-	4.29	-.48	-.67
4. Parent Language: Quantity	.15	.02	-	-	-	1.52	1.66
5. Speed	**.15**	**0**	-	-	-	-	2.39

Note: $^{\dagger}p<.10$, $^{*}p<.05$, $^{**}p< .01$, $^{***}p< .001$

Supplemental Table 8

Multiple regressions and change in R^2 for accuracy and speed of intersensory matching for social events, parent language input (both quantity and quality), and maternal education at 6 months predicting child speech production at 36 months (N = 103). The unique variance (ΔR^2) for each predictor (when holding all other predictors constant) is presented in Step 5 of each model.

	36-Month Child Speech Quantity							**36-Month Child Speech Quality**						
	Variance		**beta**					**Variance**		**beta**				
Steps and predictors: 6-Month	Total R^2	ΔR^2	Step 1	Step 2	Step 3	Step 4	Step 5	Total R^2	ΔR^2	Step 1	Step 2	Step 3	Step 4	Step 5
Model 1														
1. Maternal Education	.01	.01	.73	.53	.56	.61	.63	.04	.04	.52*	.46†	.47†	.48†	.49*
2. Parent Language: Quality	.01	.01	-	.18	-.07	-.05	-.04	.05	.01	-	.05	.02	.02	.03
3. Parent Language: Quantity	.02	.01	-	-	.09	.07	.06	.05	0	-	-	.01	.01	.004
4. Speed	.02	0	-	-	-	.50	.56	.05	0	-	-	-	.19	.23
5. Accuracy	**.02**	**0**	-	-	-	-	.15	**.08**	**.03**	-	-	-	-	.12
Model 2														
1. Parent Language: Quality	.01	.01	.27	.04	.08	.08	-.04	.02	.02	.10	.08	.08	.09	.03
2. Parent Language: Quantity	.02	.01	-	.08	.06	.05	.06	.02	0	-	.01	.004	0	.004
3. Speed	.02	0	-	-	.42	.48	.56	.02	0	-	-	.11	.16	.23
4. Accuracy	.02	0	-	-	-	.14	.15	.04	.02	-	-	-	.12	.12
5. Maternal Education	**.02**	**0**	-	-	-	-	.63	**.08**	**.04***	-	-	-	-	.49*
Model 3														
1. Parent Language: Quantity	.02	.02	.09	.12	.07	.05	.06	.02	.02	.02	.02	.02	.01	.004
2. Speed	.02	0	-	.02	.47	.56	.56	.02	0	-	.10	.14	.23	.23
3. Accuracy	.02	0	-	-	.14	.15	.15	.03	.01	-	-	.11	.12	.12
5. Maternal Education	.02	0	-	-	-	.61	.63	.08	.05*	-	-	-	.51*	.49*
5. Parent Language: Quality	**.02**	**0**	-	-	-	-	-.04	**.08**	**0**	-	-	-	-	.03
Model 4														
1. Speed	.001	.001	.56	.62	.71	.62	.56	.001	.001	.14	.18	.26	.24	.23
2. Accuracy	.004	.003	-	.15	.16	.16	.15	.02	.019	-	.12	.12	.12	.12
3. Maternal Education	.01	.006	-	-	.76	.61	.63	.07	.05*	-	-	.54*	.49*	.49*
4. Parent Language: Quality	.02	.01	-	-	-	.14	-.04	.08	.01	-	-	-	.05	.03

5. Parent Language: Quantity	**.02**	**0**	-	-	-	-	.06	**.08**	**0**	-	-	-	-	.004
Model 5														
1. Accuracy	.003	.003	.15	.05	.15	.14	.15	.02	.02	.12	.12	.12	.12	.12
2. Maternal Education	.01	.007	-	.08	.53	.57	.63	.07	.05*	-	.53*	.47†	.47†	.49*
3. Parent Language: Quality	.02	.01	-	-	.18	-.06	-.04	.08	.01	-	-	.05	.03	.03
4. Parent Language: Quantity	.02	0	-	-	-	.08	.06	.08	0	-	-	-	.01	.004
5. Speed	**.02**	**0**	-	-	-	-	.56	**.08**	**0**	-	-	-	-	.23

Note: †p<.10, *p<.05, **p< .01, ***p< .001

Supplemental Table 9

Multiple regressions and change in R^2 for accuracy and speed of intersensory matching for social events, parent language input (both quantity and quality), and maternal education at 6 months predicting child vocabulary size (expressive, receptive) at 36 months (N = 103). The unique variance (ΔR^2) for each predictor (when holding all other predictors constant) is presented in Step 5 of each model.

	36-Month Expressive Vocabulary							**36-Month Receptive Vocabulary**						
	Variance		**beta**					**Variance**		**beta**				
Steps and predictors: **6-Month**	Total R^2	ΔR^2	Step 1	Step 2	Step 3	Step 4	Step 5	Total R^2	ΔR^2	Step 1	Step 2	Step 3	Step 4	Step 5
Model 1														
1. Maternal Education	.20**	.20**	4.80***	4.49***	4.26**	4.34***	4.48***	.14	.14	4.01**	3.68*	3.72*	3.78*	4.02**
2. Parent Language: Quality	.21*	.01	-	.24	.94†	.97*	1.03*	.15	.01	-	.27	.11	.12	.16
3. Parent Language: Quantity	.24**	.03*	-	-	-.22†	-.24*	-.27*	.15	0	-	-	.05	.04	.02
4. Speed	.24**	0	-	-	-	3.03	3.64	.15	0	-	-	-	1.56	2.17
5. Accuracy	**.32*** **	**.08***	-	-	-	-	.86**	**.30****	**.15*****	-	-	-	-	1.33***
Model 2														
1. Parent Language: Quality	.04	.04	.57	1.41*	1.45**	1.53**	1.03*	.04	.04	.59	.56	.57	.62	.16
2. Parent Language: Quantity	.07	.03*	-	-.28†	-.30*	-.32*	-.27*	.04	0	-	.01	.01	-.02	.02
3. Speed	.09	.02	-	-	2.93	3.49	3.64	.04	0	-	-	1.25	1.81	2.17
4. Accuracy	.15*	.06*	-	-	-	.80*	.86**	.17*	.13**	-	-	-	1.25**	1.33***
5. Maternal Education	**.32*** **	**.17*****	-	-	-	-	4.48***	**.30****	**.13****	-	-	-	-	4.02**
Model 3														
1. Parent Language: Quantity	.001	.001	.03	.02	.01	-.05	-.27*	.03	.03	.13	.13	.11	.05	.02
2. Speed	.02	.019	-	2.54	3.03	3.48	3.64	.03	0	-	1.22	1.75	2.19	2.17
3. Accuracy	.06	.04*	-	-	.73*	.82	.86**	.14*	.11**	-	-	1.23**	1.33**	1.33***
5. Maternal Education	.28**	.22***	-	-	-	5.17***	4.48***	.30**	.16***	-	-	-	4.11**	4.02**
5. Parent Language: Quality	**.32*** **	**.04***	-	-	-	-	1.03*	**.30****	**0**	-	-	-	-	.16
Model 4														
1. Speed	.02	.02	2.62	3.09	3.44	3.33	3.64	.01	.01	1.53	2.04	2.43	2.29	2.17
2. Accuracy	.06	.04*	-	.73*	.81*	.81*	.86**	.12†	.11**	-	1.26**	1.33**	1.33**	1.33***
3. Maternal Education	.28**	.22***	-	-	5.02***	4.75***	4.48***	.29**	.17***	-	-	4.28**	4.01**	4.02**

4. Parent Language: Quality	.29**	.01	-	-	-	.20	1.03*	.30**	.01	-	-	-	.21	.16
5. Parent Language: Quantity	**.32***	**.03***	-	-	-	-	-.27*	**.30****	**0**	-	-	-	-	.02
Model 5														
1. Accuracy	.04	.04	.67*	.75*	.74*	.79*	.86**	.12†	.12†	1.23**	1.30**	1.30**	1.30**	1.33***
2. Maternal Education	.26**	.22***	-	4.93***	4.62***	4.37***	4.48***	.29**	.17***	-	4.21**	3.91**	3.94**	4.02**
3. Parent Language: Quality	.27*	.01	-	-	.24	.99*	1.03*	.29**	0	-	-	.24	.15	.16
4. Parent Language: Quantity	.29**	.02	-	-	-	-.24*	-.27*	.30**	.01	-	-	-	.03	.02
5. Speed	**.32***	**.03***	-	-	-	-	3.64	**.30****	**0**	-	-	-	-	2.17

Note: †p<.10, *p<.05, **p< .01, ***p< .001

Supplemental Table 10

Multiple regressions and change in R^2 for accuracy and speed of intersensory matching for social events at 6 months, quantity and quality of parent language input at 18 months, and maternal education in predicting quantity and quality of child speech production at 18 months (N = 103).

	Outcomes: 18-Month Child Speech Production													
	Child Speech Quantity							**Child Speech Quality**						
	Variance		**beta**					**Variance**		**beta**				
Steps and predictors	Total R^2	ΔR^2	Step 1	Step 2	Step 3	Step 4	Step 5	Total R^2	ΔR^2	Step 1	Step 2	Step 3	Step 4	Step 5
Model 1														
1. Maternal Education	.05	.05	.27	.19	.20	.21	.15	.08	.08	.14*	.11†	.11†	.12*	.09†
2. Parent Language: Quality	.09	.04†	-	.08†	.05	.04	.07	.13	.05	-	.04	.04	.04	.05
3. Parent Language: Quantity	.09	0	-	-	.01	.01	.01	.13	0	-	-	-.001	-.001	-.001
4. Speed of Selection	.11	.02	-	-	-	.29	.36	.15	.02	-	-	-	.16	.20*
5. Intersensory Matching	.24*	.13***	-	-	-	-	.15***	.32***	.17***	-	-	-	-	.07***
Model 2														
1. Parent Language: Quality	.07	.07*	.10*	.08	.08	.09	.07	.09	.09*	.05†	.06	.05	.06	.05
2. Parent Language: Quantity	.07	0	-	.004	.01	.01	.01	.09	0	-	-.003	-.002	-.003	-.001
3. Speed of Selection	.08	.01	-	-	.26	.35†	.36	.11	.02	-	-	.15	.19†	.20*
4. Intersensory Matching	.23*	.15***	-	-	-	.16***	.15***	.30**	.19***	-	-	-	.08***	.07***
5. Maternal Education	.24*	.01	-	-	-	-	.15	.32***	.02†	-	-	-	-	.09†
Model 3														
1. Parent Language: Quantity	.06*	.06*	.02*	.02*	.02*	.02*	.01	.05†	.05†	.01	.007	.01*	.01†	-.001
2. Speed of Selection	.07	.01	-	.28	.37†	.39†	.36	.08	.03	-	.16	.20†	.21*	.20*
3. Intersensory Matching	.21**	.14***	-	-	.16***	.15***	.15***	.26**	.18***	-	-	.07***	.07***	.07***
5. Maternal Education	.23*	.02	-	-	-	.18	.15	.29**	.03*	-	-	-	.11*	.09†
5. Parent Language: Quality	.24*	.01	-	-	-	-	.07	.32***	.03	-	-	-	-	.05
Model 4														
1. Speed of Selection	.02	.02	.31	.40†	.42†	.36†	.36	.03	.03	.17	.21†	.22*	.20*	.20*
2. Intersensory Matching	.14†	.12**	-	.14**	.13**	.16***	.15***	.19**	.16***	-	.07***	.07***	.07***	.07***
3. Maternal Education	.17†	.03	-	-	.24	.14	.15	.25**	.06**	-	-	.13*	.09†	.09†

4. Parent Language: Quality	.23*	.06*	-	-	-	.10*	.07	.32**	.07*	-	-	-	.04†	.05
5. Parent Language: Quantity	.24*	.01	-	-	-	-	.01	.32***	0	-	-	-	-	-.001
Model 5														
1. Intersensory Matching	.11**	.11**	.14**	.13**	.15***	.15***	.15***	.14***	.14***	.38***	.06***	.07***	.07***	.07***
2. Maternal Education	.14	.03	-	.23	.12	.13	.15	.20*	.06*	-	.13*	.08	.08	.09†
3. Parent Language: Quality	.21*	.07*	-	-	.11*	.08	.07	.28**	.08*	-	-	.05†	.05	.05
4. Parent Language: Quantity	.22*	.01	-	-	-	.01	.01	.28**	0	-	-	-	-.002	-.001
5. Speed of Selection	.24*	.02†	-	-	-	-	.36	.32***	.04*	-	-	-	-	.20*

Note: †p<.10, *p<.05, **p< .01, ***p< .001

Supplemental Table 11

Multiple regressions and change in R^2 for accuracy and speed of intersensory matching for social events at 6 months, quantity and quality parent language input at 18 months, and maternal education in predicting expressive and receptive child vocabulary size at 18 months (N = 103).

	Outcomes: 18-Month Child Vocabulary Size													
	Expressive Vocabulary							**Receptive Vocabulary**						
	Variance		**beta**					**Variance**		**beta**				
Steps and predictors	Total R^2	ΔR^2	Step 1	Step 2	Step 3	Step 4	Step 5	Total R^2	ΔR^2	Step 1	Step 2	Step 3	Step 4	Step 5
Model 1														
1. Maternal Education	.03	.03	9.48	5.87	6.57	6.68	4.54	.003	.003	5.31	-3.82	-3.93	-3.14	-3.04
2. Parent Language: Quality	.07	.04	-	3.41	.74	.61	.72	.07	.067*	-	8.46*	8.20	8.56	8.46
3. Parent Language: Quantity	.08	.01	-	-	.62	.61	.61	.07	0	-	-	.06	.09	.09
4. Speed of Selection	.08	0	-	-	-	6.19	14.92	.09	.02	-	-	-	-26.74	-27.73
5. Intersensory Matching	.21*	.13**	-	-	-	-	6.95*	.09	0	-	-	-	-	-.71
Model 2														
1. Parent Language: Quality	.06	.06	3.80	1.51	1.42	1.12	.72	.07	.07**	8.23*	7.72	8.17	8.23	8.46
2. Parent Language: Quantity	.06	0	-	.54	.53	.58	.61	.07	0	-	.12	.14	.11	.09
3. Speed of Selection	.06	0	-	-	6.70	15.59	14.92	.09	.02	-	-	-26.51	-27.94	-27.73
4. Intersensory Matching	.20†	.14***	-	-	-	7.09*	6.95*	.09	0	-	-	-	-1.07	-.71
5. Maternal Education	.21*	.01	-	-	-	-	4.54	.09	0	-	-	-	-	-3.04
Model 3														
1. Parent Language: Quantity	.06	.06	.81	.78	.77	.73	.61	.06*	.06*	1.51*	1.59*	1.58*	1.55*	.09
2. Speed of Selection	.06	0	-	7.19	15.94	15.30	14.92	.08	.02	-	-24.27	-25.19	-25.26	-27.73
3. Intersensory Matching	.20†	.14**	-	-	7.04*	6.84*	6.95*	.08	0	-	-	-.70	-.70	-.71
5. Maternal Education	.20*	0	-	-	-	5.11	4.54	.08	0	-	-	-	1.29	-3.04
5. Parent Language: Quality	.21*	.01	-	-	-	-	.72	.09	.01	-	-	-	-	8.46
Model 4														
1. Speed of Selection	.01	.01	10.01	18.72	17.63	15.30	14.92	.01	.01	-18.14	-18.38	-19.67	-27.38	-27.73
2. Intersensory Matching	.16	.15**	-	6.93*	6.60*	6.96*	6.95*	.01	0	-	-.56	-.86	-.74	-.71
3. Maternal Education	.17	.01	-	-	7.57	3.84	4.54	.02	.01	-	-	6.32	-2.97	-3.04

4. Parent Language: Quality	.20†	.03	-	-	-	3.38	.72	.09	.07*	-	-	-	8.79*	8.46
5. Parent Language: Quantity	.21*	.01	-	-	-	-	.61	.09	0	-	-	-	-	.09
Model 5														
1. Intersensory Matching	.11**	.11**	5.85*	5.57*	.34**	6.08*		0	0	.48	.27	.86	.92	-.71
2. Maternal Education	.12	.01	-	8.12	.08	4.80		.003	.003	-	5.14	-4.14	-4.26	-3.04
3. Parent Language: Quality	.17†	.05	-	-	.22	.88		.06	.057*	-	-	8.32*	8.00	8.46
4. Parent Language: Quantity	.18*	.01	-	-	-	.63		.07	.01	-	-	-	.08	.09
5. Speed of Selection	.21*	.03	-	-	-	-		.09	.02	-	-	-	-	-27.73

Note: $^{\dagger}p<.10$, $^{*}p<.05$, $^{**}p< .01$, $^{***}p< .001$

Supplemental Table 12

Multiple regressions and change in R^2 for accuracy and speed of intersensory matching for social events at 6 months, quantity and quality of parent language input at 24 months, and maternal education in predicting quantity and quality of child speech production at 24 months (N = 103).

	Outcomes: 24-Month Child Speech Production													
	Child Speech Quantity							**Child Speech Quality**						
	Variance		**beta**					**Variance**		**beta**				
Steps and predictors	Total R^2	ΔR^2	Step 1	Step 2	Step 3	Step 4	Step 5	Total R^2	ΔR^2	Step 1	Step 2	Step 3	Step 4	Step 5
Model 1														
1. Maternal Education	.11**	.11**	1.18*	.90†	.90†	.92†	.92*	.21***	.21***	.67***	.52**	.52**	.52**	.52***
2. Parent Language: Quality	.15	.04†	-	.20	.20	.19	.14	.27**	.16**	-	.11*	.11	.10	.09
3. Parent Language: Quantity	.15	0	-	-	-.001	.003	.01	.28**	.01	-	-	0	.002	.003
4. Speed of Selection	.15†	0	-	-	-	.71	.91	.29**	.01	-	-	-	.38	.45
5. Intersensory Matching	.31**	.16***	-	-	-	-	.47***	.41***	.12***	-	-	-	-	.16***
Model 2														
1. Parent Language: Quality	.08*	.08*	.27*	.30	.29	.24	.14	.16***	.16***	.15**	.16†	.16†	.15†	.09
2. Parent Language: Quantity	.08	0	-	-.01	-.004	.001	.01	.16†	0	-	-.004	-.002	-.001	.003
3. Speed of Selection	.09	.01	-	-	.68	.87	.91	.17*	.01	-	-	.36	.43	.45
4. Intersensory Matching	.24*	.15***	-	-	-	.48***	.47***	.29**	.12***	-	-	-	.17**	.16***
5. Maternal Education	.31**	.07*	-	-	-	-	.92*	.41***	.12***	-	-	-	-	.52***
Model 3														
1. Parent Language: Quantity	.05*	.05*	.06*	.06*	.06*	.04†	.01	.10**	.10**	.03**	.03**	.03**	.02**	.003
2. Speed of Selection	.07	.02	-	.78	.96	.97	.91	.12†	.02	-	.40	.47†	.48†	.45
3. Intersensory Matching	.22*	.15***	-	-	.49***	.48***	.47***	.24*	.12***	-	-	.17**	.17***	.16***
5. Maternal Education	.30**	.08*	-	-	-	.99*	.92*	.39***	.15***	-	-	-	.58***	.52***
5. Parent Language: Quality	.31**	.01	-	-	-	-	.14	.41***	.02	-	-	-	-	.09
Model 4														
1. Speed of Selection	.01	.01	.72	.90	.94	.92	.91	.02	.02	.37	.44	.46	.46†	.45
2. Intersensory Matching	.18*	.17***	-	.50***	.49***	.48***	.47***	.14†	.12***	-	.17**	.17***	.16***	.16***
3. Maternal Education	.28**	.10**	-	-	1.15**	.91*	.92*	.35***	.21***	-	-	.66***	.52***	.52***

4. Parent Language: Quality	.31**	.03†	-	-	-	.17†	.14	.41***	.13**	-	-	-	.10**	.09
5. Parent Language: Quantity	.31**	0	-	-	-	-	.01	.41***	0	-	-	-	-	.003
Model 5														
1. Intersensory Matching	.17*	.17***	.49***	.48***	.46***	.46***	.47***	.13**	.13**	.17**	.17***	.16***	.16***	.16***
2. Maternal Education	.27**	.10**	-	1.14**	.89*	.90*	.92*	.33***	.20***	-	.65***	.51**	.51**	.52***
3. Parent Language: Quality	.30**	.03†	-	-	.18†	.16	.14	.39***	.12**	-	-	.10**	.10	.09
4. Parent Language: Quantity	.30**	0	-	-	-	.004	.01	.39***	0	-	-	-	.001	.003
5. Speed of Selection	.31**	.01	-	-	-	-	.91	.41***	.02†	-	-	-	-	.45

Note: †p<.10, *p<.05, **p< .01, ***p< .001

Supplemental Table 13

Multiple regressions and change in R^2 for accuracy and speed of intersensory matching for social events at 6 months, quantity and quality parent language input at 24 months, and maternal education in predicting expressive child vocabulary size at 24 months (N = 103).

	24-Month Child Vocabulary Size Expressive Vocabulary						
	Variance		beta				
Steps and predictors	Total R^2	ΔR^2	Step 1	Step 2	Step 3	Step 4	Step 5
Model 1							
1. Maternal Education	.06†	.06†	30.59†	15.48	24.35	24.61	26.69
2. Parent Language: Quality	.10	.04*	-	7.92†	-10.20	-10.47	-12.10
3. Parent Language: Quantity	.25*	.15**	-	-	5.65**	5.72**	5.95**
4. Speed of Selection	.25*	0	-	-	-	-3.09	2.66
5. Intersensory Matching	.32**	.07*	-	-	-	-	10.26*
Model 2							
1. Parent Language: Quality	.10**	.10**	9.99**	-5.18	-5.40	-6.73	-12.10
2. Parent Language: Quantity	.21*	.11**	-	4.91**	4.97**	5.18**	5.95**
3. Speed of Selection	.21*	0	-	-	-6.06	-.71	2.66
4. Intersensory Matching	.27**	.06*	-	-		9.56†	10.26*
5. Maternal Education	.32**	.05	-	-		-	26.69
Model 3							
1. Parent Language: Quantity	.19***	.19***	3.65**	3.66**	3.58**	3.24**	5.95**
2. Speed of Selection	.19*	0	-	-6.13	-.73	.61	2.66
3. Intersensory Matching	.24**	.05*	-	-	9.06*	9.36*	10.26*
5. Maternal Education	.26**	.02	-	-	-	15.77	26.69
5. Parent Language: Quality	.32**	.06	-	-	-	-	-12.10
Model 4							
1. Speed of Selection	.001	.001	-6.43	-.57	2.83	1.52	2.66
2. Intersensory Matching	.06	.059*	-	9.86*	10.0*	9.32*	10.26*
3. Maternal Education	.13	.07†	-	-	30.72†	16.65	26.69
4. Parent Language: Quality	.16†	.03†	-	-	-	7.36†	-12.10
5. Parent Language: Quantity	.32**	.16**	-	-	-	-	5.95**
Model 5							
1. Intersensory Matching	.06*	.06*	9.87*	9.96*	9.30*	10.24*	10.26*
2. Maternal Education	.13	.07†	-	30.65†	16.60	26.26	26.69
3. Parent Language: Quality	.16†	.03†	-	-	7.37†	-11.82	-12.10
4. Parent Language: Quantity	.32**	.16**	-	-	-	5.88**	5.95**
5. Speed of Selection	.32**	0	-	-	-	-	2.66

Note: $^{\dagger}p<.10$, $^{*}p<.05$, $^{**}p< .01$, $^{***}p< .001$

Supplemental Table 14

Multiple regressions and change in R^2 for accuracy and speed of intersensory matching for social events at 6 months, quantity and quality of parent language input at 36 months, and maternal education in predicting quantity and quality of child speech production at 36 months (N = 103).

	Outcomes: 36-Month Child Speech Production													
	Child Speech Quantity							**Child Speech Quality**						
	Variance		**beta**					**Variance**		**beta**				
Steps and predictors	Total R^2	ΔR^2	Step 1	Step 2	Step 3	Step 4	Step 5	Total R^2	ΔR^2	Step 1	Step 2	Step 3	Step 4	Step 5
Model 1														
1. Maternal Education	.01	.01	.73	.06	.15	.18	.20	.04*	.04*	.52*	.23	.24	.24	.26
2. Parent Language: Quality	.04	.03†	-	.54†	.28	.27	.24	.12†	.08**	-	.24*	.22†	.22†	.20†
3. Parent Language: Quantity	.04	0	-	-	.06	.06	.07	.12†	0	-	-	.004	.004	.01
4. Speed of Selection	.04	0	-	-	-	.35	.41	.12†	0	-	-	-	.07	.12
5. Intersensory Matching	.05	.01	-	-	-	-	.14	.14†	.02	-	-	-	-	.11
Model 2														
1. Parent Language: Quality	.04†	.04†	.54†	.11	.30	.28	.24	.11**	.11**	.26**	.26**	.25*	.24*	.20†
2. Parent Language: Quantity	.04	0	-	.11	.06	.07	.07	.11	0	-	.001	.001	.004	.01
3. Speed of Selection	.04	0	-	-	.36	.41	.41	.11	0	-	-	.03	.07	.12
4. Intersensory Matching	.05	.01	-	-	-	.14	.14	.13†	.02	-	-	-	.11	.11
5. Maternal Education	.05	0	-	-	-	-	.20	.14†	.01	-	-	-	-	.26
Model 3														
1. Parent Language: Quantity	.04	.04	.11	.11	.11	.10	.07	.07*	.07*	.04*	.04*	.04*	.04†	.01
2. Speed of Selection	.04	0	-	.45	.50	.54	.41	.07	0	-	.10	.14	.21	.12
3. Intersensory Matching	.04	0	-	-	.16	.17	.14	.09	.02	-	-	.12	.12	.11
5. Maternal Education	.05	.01	-	-	-	.38	.20	.12	.03†	-	-	-	.41†	.26
5. Parent Language: Quality	.05	0	-	-	-	-	.24	.14†	.02†	-	-	-	-	.20†
Model 4														
1. Speed of Selection	.001	.001	.56	.62	.71	.34	.41	.001	.001	.14	.18	.26	.11	.12
2. Intersensory Matching	.004	.003	-	.15	.16	.13	.14	.02	.019	-	.12	.12	.11	.11
3. Maternal Education	.01	.006	-	-	.76	.09	.20	.07	.05*	-	-	.54*	.25	.26

4. Parent Language: Quality	.04	.03†	-	-	-	.53	.24	.14†	.07**	-	-	-	.23**	.20†
5. Parent Language: Quantity	.05	.01	-	-	-	-	.07	.14†	0	-	-	-	-	.01
Model 5														
1. Intersensory Matching	.003	.003	.15	.05	.12	.14	.14	.02	.02	.12	.12	.11	.11	.11
2. Maternal Education	.01	.007	-	.08	.07	.18	.20	.07	.05*	-	.53*	.24	.25	.26
3. Parent Language: Quality	.04	.03†	-	-	.53	.26	.24	.14†	.07**	-	-	.23**	.20†	.20†
4. Parent Language: Quantity	.05	.01	-	-	-	.07	.07	.14†	0	-	-	-	.01	.01
5. Speed of Selection	.05	0	-	-	-	-	.41	.14†	0	-	-	-	-	.12

Note: $^{\dagger}p<.10$, $^{*}p<.05$, $^{**}p< .01$, $^{***}p< .001$

Supplemental Table 15

Multiple regressions and change in R^2 for accuracy and speed of intersensory matching for social events at 6 months, quantity and quality parent language input at 36 months, and maternal education in predicting expressive and receptive child vocabulary size at 36 months (N = 103).

	Outcomes: 36-Month Child Vocabulary Size													
	Expressive Vocabulary							**Receptive Vocabulary**						
	Variance		**beta**					**Variance**		**beta**				
Steps and predictors	Total R^2	ΔR^2	Step 1	Step 2	Step 3	Step 4	Step 5	Total R^2	ΔR^2	Step 1	Step 2	Step 3	Step 4	Step 5
Model 1														
1. Maternal Education	$.20^{***}$	$.20^{***}$	4.80^{***}	3.48^{**}	3.17^{**}	3.27^{**}	3.53^{**}	$.14^{***}$	$.14^{***}$	4.01^{**}	$2.47^{\dagger}$	2.33	2.39	2.82^{*}
2. Parent Language: Quality	$.27^{**}$	$.07^{*}$	-	1.01^{*}	1.69^{**}	1.63^{**}	1.47^{**}	$.24^{**}$	$.10^{**}$	-	1.23^{**}	1.55^{*}	1.50^{*}	1.21^{*}
3. Parent Language: Quantity	$.29^{**}$	$.02^{\dagger}$	-	-	$-.16^{\dagger}$	$-.16^{\dagger}$	$-.13^{\dagger}$	$.25^{**}$	.01	-	-	-.08	-.07	-.03
4. Speed of Selection	$.30^{**}$	.01	-	-	-	2.32	2.83	$.25^{**}$	0	-	-	-	1.35	1.97
5. Intersensory Matching	$.35^{***}$	$.05^{*}$	-	-	-	-	$.68^{*}$	$.37^{***}$	$.12^{***}$	-	-	-	-	1.24^{***}
Model 2														
1. Parent Language: Quality	$.17^{***}$	$.17^{***}$	1.38^{**}	2.18^{***}	2.15^{***}	2.05^{***}	1.47^{**}	$.20^{***}$	$.20^{***}$	1.54^{***}	1.98^{**}	1.96^{**}	1.75^{**}	1.21^{*}
2. Parent Language: Quantity	$.21^{*}$	$.04^{*}$	-	$-.20^{*}$	$-.20^{*}$	$-.18^{*}$	$-.13^{\dagger}$	$.21^{*}$	.01	-	-.11	-.11	-.08	-.03
3. Speed of Selection	$.22^{*}$	.01	-	-	2.12	2.52	2.83	$.22^{*}$	.01	-	-	1.16	1.66	1.97
4. Intersensory Matching	$.24^{*}$	$.02^{*}$	-	-	-	$.58^{\dagger}$	$.68^{*}$	$.31^{**}$	$.09^{***}$	-	-	-	1.14^{***}	1.24^{***}
5. Maternal Education	$.35^{***}$	$.11^{**}$	-	-	-	-	3.53^{**}	$.37^{***}$	$.06^{*}$	-	-	-	-	2.82^{*}
Model 3														
1. Parent Language: Quantity	$.03^{\dagger}$	$.03^{\dagger}$	.12	.12	.12	.07	$-.13^{\dagger}$	$.08^{*}$	$.08^{*}$	$.19^{*}$	$.19^{*}$	$.19^{*}$	$.15^{\dagger}$	-.03
2. Speed of Selection	.05	.02	-	2.64	3.12	3.42	2.83	.09	.01	-	1.62	2.13	2.44	1.97
3. Intersensory Matching	.10	$.05^{*}$	-	-	$.74^{*}$	$.82^{*}$	$.68^{*}$	$.21^{*}$	$.12^{***}$	-	-	1.28^{***}	1.35^{***}	1.24^{***}
5. Maternal Education	$.29^{**}$	$.19^{***}$	-	-	-	4.76^{***}	3.53^{**}	$.34^{***}$	$.13^{**}$	-	-	-	3.80^{**}	2.82^{*}
5. Parent Language: Quality	$.35^{***}$	$.06^{**}$	-	-	-	-	1.47^{**}	$.37^{***}$	$.03^{*}$	-	-	-	-	1.21^{*}
Model 4														
1. Speed of Selection	.02	.02	2.62	3.09	3.44	3.06	2.83	.01	.01	1.53	2.04	2.43	2.02	1.97
2. Intersensory Matching	.06	$.04^{*}$	-	$.73^{*}$	$.81^{*}$	$.74^{*}$	$.68^{*}$	$.12^{\dagger}$	$.11^{***}$	-	1.26^{**}	1.33^{**}	1.25^{***}	1.24^{***}
3. Maternal Education	$.28^{**}$	$.22^{***}$	-	-	5.02^{***}	3.80^{***}	3.53^{**}	$.29^{**}$	$.17^{***}$	-	-	4.28^{**}	2.86^{*}	2.82^{*}

4. Parent Language: Quality	.34**	.06*	-	-	-	.91*	1.47**	.37***	.08**	-	-	-	1.10**	1.21*
5. Parent Language: Quantity	.35***	0	-	-	-	-	-.13†	.37***	0	-	-	-	-	-.03
Model 5														
1. Intersensory Matching	.04*	.04*	.67*	.75*	.67†	.62†	.68*	.12***	.12***	1.23**	1.30**	1.22***	1.21***	1.24***
2. Maternal Education	.26**	.22***	-	4.93***	3.66**	3.38**	3.53**	.29**	.17***	-	4.21**	2.78*	2.72*	2.82*
3. Parent Language: Quality	.32**	.06**	-	-	.95*	1.56**	1.47**	.37***	.08**	-	-	1.13**	1.28*	1.21*
4. Parent Language: Quantity	.33**	.01	-	-	-	-.15†	-.13†	.37***	0	-	-	-	-.04	-.03
5. Speed of Selection	.35***	.02	-	-	-	-	2.83	.37***	0	-	-	-	-	1.97

Note: †p<.10, *p<.05, **p< .01, ***p< .001

Supplemental Table 16

Amount of unique variance accounted for by each predictor variable (accuracy and speed of intersensory matching for social events at 6 months, quantity and quality of parent language input at 18, 24, or 36 months, and maternal education) in predicting child language outcomes at 18, 24, or 36 months (N = 103).

	18-Month Language Outcomes			
Predictors	Quantity	Quality	Expressive	Receptive
Total Variance	.24*	.32***	.21*	.09
Unique Variance				
6-Month Intersensory Matching				
Accuracy	.13***	.17***	.13**	0
Speed	.02†	.04*	.03	.02
18-Month Parent Language Input				
Quantity	.01	0	.01	0
Quality	.01	.03	.01	.01
Maternal Education	.01	.02†	.01	0

	24-Month Language Outcomes		
	Quantity	Quality	Expressive
Total Variance	.31**	.41***	.32**
Unique Variance			
6-Month Intersensory Matching			
Accuracy	.16***	.12***	.07*
Speed	.01	.02†	0
24-Month Parent Language Input			
Quantity	0	0	.16**
Quality	.01	.02	.06
Maternal Education	.07*	.12***	.05

	36-Month Language Outcomes			
	Quantity	Quality	Expressive	Receptive
Total Variance	.05	.14†	.35***	.37***
Unique Variance				
6-Month Intersensory Matching				
Accuracy	.01	.02	.05*	.12***
Speed	0	0	.02	0
36-Month Parent Language Input				
Quantity	.01	0	0	0
Quality	0	.02†	.06**	.03*
Maternal Education	0	.01	.11**	.06*

Note: †p<.10, *p<.05, **p< .01, ***p< .001

VITA

ELIZABETH V. EDGAR

2016	B.S., Human Development and Family Studies Pennsylvania State University Scranton, PA
2019	M.S., Psychology Florida International University Miami, FL
2020 – 2021	Doctoral Candidate Psychology Florida International University Miami, FL

PUBLICATIONS AND SELECT PRESENTATIONS

Edgar, E. V., Todd, J. T., & Bahrick, L. E. (2019, March). *Developmental Pathways to Language: Intersensory Processing and Child Gesture Production*. Poster presented at the International Convention of Psychological Science, Paris, France.

Edgar, E. V., Todd, J. T., & Bahrick, L. E. (2020, July). *Multisensory Attention Skills and Parent Language Input Predict Children's Vocabulary Across 12 to 18 Months of Age*. Virtual poster presentation, Biennial Meeting for the International Congress of Infant Studies.

Edgar, E. V., Todd, J. T., & Bahrick, L. E. (2021). Intersensory Matching of Faces and Voices in Infancy Predicts Language Outcomes in Young Children. [Manuscript under review]. Department of Psychology, Florida International University.

Edgar, E. V., Todd, J. T., & Bahrick, L. E. (2021.). Intersensory Processing of Social Events at 6 Months Predicts Language Outcomes at 18, 24, and 36 Months of Age. [Manuscript under review]. Department of Psychology, Florida International University.

Edgar, E. V., Todd, J. T., Eschman, B., & Bahrick, L. E. (2021). Effects of English Versus Spanish Language Exposure on Tests of Multisensory Attention Skills Across 3 to 36 Months of Age. [Manuscript in preparation]. Department of Psychology, Florida International University.

Edgar, E. V., Todd, J. T., McNew, M. E., & Bahrick, L. E. (2018, October). *Relations*

Among Intersensory Processing, Social Competence, and Vocabulary Size in Infancy. Poster presented at the International Society for Developmental Psychobiology, San Diego, CA.

Eschman, B., Todd, J. T., Sarafraz, A., Edgar, E. V., Petrulla, V., McNew, M., & Bahrick, L. E. (2021). Data Collection During a Pandemic: A new approach for quantifying looking time data in infants and young children [Manuscript under review]. Department of Psychology, Florida International University.