6-19-2023

A comparison of voice and gesture across the first two years of life

Megan Burkhardt-Reed

A COMPARISON OF VOICE AND GESTURE ACROSS THE FIRST TWO YEARS OF LIFE

by

Megan M. Burkhardt-Reed

A Dissertation

Submitted in Partial Fulfillment of the

Requirements for the Degree of

Doctor of Philosophy

Major: Communication Sciences and Disorders

University of Memphis

May 2023

Dedication

To my mentor, Dr. D. Kimbrough Oller, I could not have done this without all of your encouragement and constructive criticisms that both opened my mind and improved the quality of my work. I will be forever grateful for your guidance and unwavering support. For my family, thank you for always challenging me to meet my full potential and your unconditional love, support, and encouragement throughout my life. To my dear husband Joseph, I know that this dream of mine would have never come to fruition without your push and your support despite all life has thrown at us.

Acknowledgements

This work was supported by the Plough Foundation and grants from the National Institute of Deafness and Communication Disorders of the National Institutes of Health, R01DC011027 and R01DC006099, awarded to D. Kimbrough Oller. Scholarly travel funding was also provided by the Institute of Intelligent Systems Student Organization of the University of Memphis.

Preface

This dissertation was written using the journal-ready article format. Chapter 2 of this dissertation was published as a journal article in 2021 in *Infant Behavior and Development*. Its authors are Megan M. Burkhardt-Reed, Helen L. Long, Dale D. Bowman, Edina R. Bene, and D. Kimbrough Oller. Chapter 3 is to be submitted as a manuscript to *PLoS ONE* for publication. Its authors are Megan M. Burkhardt-Reed and D. Kimbrough Oller.

Abstract

Burkhardt-Reed, Megan M. PhD. The University of Memphis. May 2023. A comparison of voice and gesture across the first two years of life. Major Professor: D. Kimbrough Oller, PhD.

This dissertation compared gestural and vocal communication in the development of language in early infancy/childhood. The work also has implications regarding the evolution of language. Since language is primarily vocal it might be assumed vocalization is the predominant communication in infancy and that the evolution of language also depended primarily on the evolution of vocal capabilities. But the primary literature actually favors primarily gestural language origins. The present work contradicts the primary literature.

Study 1 examined rates of gesture and speech-like vocalizations, or "protophones", in the first year of life. Infant protophones occurred more than 5 times more often than gestures. Gaze direction toward a possible receiver was rare for both vocalization and gesture, but vocalizations occurred more frequently with directed gaze than gestures. The results thus contradict the widespread belief that early language is founded primarily in gesture, and the gaze directivity data add to the contradiction. Gesture is useless as communication if no one is looking. Yet vocalization, which can communicate without listeners watching, was significantly more often accompanied by gaze directed to caregivers than gesture was. It appeared, therefore, that a greater proportion of vocalizations than gestures in the first year may have been intended as communications.

Study 2 evaluated how often children produced gestures and vocalizations (i.e., protophones and words) in the second year of life (at 13, 16 and 20 months). As with Study 1, the results suggested vocalization played a much more important role in language learning than

gesture. Gestural activity occurred much more often in the second year than in the first, but vocalization still exceeded gestural acts by more than a factor of two. More importantly, the vast majority of gestures were confined to Universal acts that are not symbolic, but rather constitute deictic indicators (pointing and reaching) that can serve no other communicative functions. In contrast, words or signs can reference abstract categories and can serve a vast array of communicative functions. Words, however, outnumbered signs by a factor greater than 11 across the data at all ages and by a factor of 21 at 20 months.

Table of Contents

List of Tables

List of Figures

1. Introduction

The origin of language has been a longstanding focus of speculation and investigation (Condillac, 1756; Hewes, 1973; Corballis, 2010). Theoretical debates about the origin of language center on whether vocalization or gesture played a more fundamental role, with prominent thinkers supporting both possibilities (Call & Tomasello, 2007; Armstrong & Wilcox, 2007; Liszkowski et al., 2011; Sterelny, 2012; Gillespie-Lynch et al., 2014; Oller et al., 2019). It has been claimed that the endogenous nature of early human infant vocal behaviors provides evidence for a foundational role of vocalization in language acquisition (Iyer et al., 2016; Oller et al., 2019, Long et al., 2020). However, the widespread claim that gesture paves the way for language development in modern human infants has furthered speculation about gestural origins (Iverson & Goldin-Meadow, 2005; Gillespie-Lynch et al., 2013). The study of modern human development can reveal not only a great deal about the origin of language, but also inform our understanding of the selection pressures that differentiated us from our ape relatives (Oller et al., 2016).

Using an evolutionary-developmental biology (evo-devo) framework (Arthur, 2021), I follow the line of thinking that presumes a natural logic of foundational stages of infant vocal development prior to the emergence of more advanced linguistic skills in humans. Any utterance in mature human language necessitates the capacity to be used as an expression that is not tied to a particular social function (Austin, 1962). That is, our system of expression must be functionally flexible (Oller et al. 2013). Any word such as “horse”, for example, must be expressible to perform any of the following illocutionary functions: naming, designating, answering, questioning, correcting, insulting, and so on. The human infant proves both inclined and able to produce several distinct speech-like vocalizations (“protophones”) freely to express positive,

negative and neutral affect on different occasions of use. This functional flexibility shows the earliest illocutionary flexibility, a capacity that is critical in language, because illocutionary flexibility is a necessary property of every element of language (every syllable, every word, every sentence, etc.). These human vocal capacities may have been naturally selected to occur early in life because of the importance of hominin infant vocalization as a signal of wellness for the altricial hominin infant (Locke, 2006).

But what about gesture? Was it similarly selected? Both vocal and gestural theories commonly reference the communicative behaviors of nonhuman primates, primarily the great apes, as evidence regarding the origin of language. Since great apes are claimed to communicate more flexibly in a visual-gestural modality than a vocal one, it has often been argued that our early hominin ancestors must have been gestural communicators (Tomasello, 2010). I reasoned that if language originated from gestural use, gestural activity should occur to a greater extent than vocal activity in early life.

The empirical goal of this dissertation is to determine the relative extent to which infants produce communicative or potentially communicative gestures and vocalizations across the first two years of life. The key to allowing sensible quantitative comparison is to determine numbers of communicative and/or potentially communicative events occurring in both infant gesture and vocalization.

Study 1 in this dissertation (Chapter 2) examined the rates of gesture and speech-like vocalizations, or "protophones", during the first year of life. If gesture occurs at a higher rate than vocalization across the first year, the results would support the idea that human communication evolved from a primarily gestural mode. This study utilized a new framework for comparison between gesture and vocalization in early communication development. The

development of a framework for optimal comparison may help to provide us with a better way to characterize early communicative acts in both domains and to quantify communicative events for comparison.

A growing body of evidence (Kersken et al., 2019) suggests vocalization occurs more often than gesture in early communication development, suggesting a more foundational role of voice. The second study in this dissertation (Chapter 3) evaluated how often children produced gestures and vocalizations (i.e., protophones and words) in the second year of life, extending the same kind of inquiry that was conducted in Study 1 and expanding the framework of description to accommodate early words and signs. Study 2 quantified proportions of vocalizations and gestures produced by children and the functions of each event. Study 2 offers perspective on the relative roles of voice and gesture in communication development throughout the second year of life and provides further speculations about the significance of the findings in thoughts about the origin of language.

2. The Origin of Language and Relative Roles of Voice and Gesture in Early Communication (Burkhardt-Reed et al., 2021)

Abstract

Both vocalization and gesture are universal modes of communication and fundamental features of language development. The gestural origins theory proposes that language evolved out of early gestural use. However, evidence reported here suggests vocalization is much more prominent in early human communication than gesture is. To our knowledge no prior research has investigated the rates of emergence of both gesture and vocalization across the first year in human infants. We evaluated the rates of gestures and speech-like vocalizations (protophones) in 10 infants at 4, 7, and 11 months of age using parent-infant laboratory recordings. We found that infant protophones outnumbered gestures substantially at all three ages, ranging from >35 times more protophones than gestures at 3 months, to >2.5 times more protophones than gestures at 11 months. The results suggest vocalization, not gesture, is the predominant mode of communication in human infants in the first year.

Introduction

Considerable controversy about the origin of language has centered on whether vocalization or gesture played a more fundamental role (Armstrong & Wilcox, 2007; Liszkowski et al., 2011; Oller et al., 2019; Sterelny, 2012). A gestural origins theory seems to have the upper hand currently (Caselli et al. 2012; Arbib et al. 2008), invoking a key argument that since great apes communicate more flexibly in the visual-gestural than the vocal modality, our hominin ancestors must have been gestural communicators (Call & Tomasello, 2007). Another argument cites reports claiming gestural communication begins earlier than vocal communication in human infants (Caselli et al, 2012).

In support of vocal origins of language, some suggest that vocal capabilities of great apes have been underestimated (Cheney & Seyfarth, 2005; Lameira, 2017) and that vocalization has the advantage of communicating in darkness or when receivers are not looking (Kendon, 2017). The massive amount of early infant vocal behavior and communication is also consistent with a primarily vocal foundation (Oller, Caskey, et al., 2019).

The present paper aims for the first time to quantify rates of speech-like vocalization and gesture across the first year to gain insight into which modality may play the more fundamental role. We also address the extent to which gestures and vocalizations are directed to potential receivers by gaze direction.

Gesture as a Foundation for Language

Although infants vocalize from birth, many believe gesture provides the first communicative opportunity (Bates, 1976; Iverson & Goldin-Meadow, 2005; Silva Lima & Cruz-Santos, 2012) and the primary driving force for the development of symbols (Bates et al., 1979; Gillespie-Lynch et al., 2013; Orr, 2018). Furthermore, research on early communicative behaviors has emphasized gestures as the first means to convey and structure communicative intent (Bates et al., 1979). Evidence has been presented to suggest that children use meaningful gestures several months before they use words (Caselli et al. 2012).

Pointing is viewed as constituting primitive deixis, a foundation for word learning (Iverson & Wozniak, 2016; Volterra et al., 2005; Tomasello et al., 2007). Caregivers have the opportunity to label objects of shared interest as infants begin to understand and to use pointing (Wu & Gros-Louis, 2015). However, pointing does not constitute naming, but instead designates entities that can be named.

Some theorists have supported the idea that language evolved from a primarily gestural mode (Arbib et al., 2008; Corballis, 2010; Hewes, 1973; Tomasello, 2010); great apes in captivity have shown both deliberate and voluntary gestures with distinctive functions (Byrne et al., 2017). Gestural symbols in human infants have been found to become less frequent than vocal symbols with age, but age-matched ape infants appear to use gestural symbols increasingly frequently across development (Gillespie-Lynch et al., 2013).

Research suggests captive apes communicate primarily through gesture (Pika et al., 2005; Pollick & De Waal, 2007; Tomasello & Zuberbühler, 2002). But unlike humans, our ape relatives do not normally assemble vocalizations into complex utterances composed of syllables, words, and sentences (Riede et al., 2005), with perhaps rare exceptions (see e.g., Clay & Zuberbühler, 2009). Gestural flexibility as opposed to vocal flexibility in great apes is supported by relatively successful sign language learning in great apes raised by humans, but almost total failure to learn spoken language in the same circumstances (Bonvillian & Patterson, 1999; Gardner & Gardner, 1969; Rivas, 2005).

Vocalization as a Foundation for Language

A variety of non-cry, speech-like vocalizations, called "protophones", are a primary means by which infants engage in communicative interaction soon after birth (Gratier et al., 2015). By ~7 months protophones come to include well-formed (or "canonical") syllables (Oller, 1980). Longitudinal investigations indicate newborn infants produce protophones from the first weeks (Koopsmans-van Beinum & Van der Stelt, 1986; Oller, 2000; Stark, 1980; Nathani et al., 2006), at much higher rates than cries in both preterm and full-term infants, with preterms producing protophones in neonatal intensive care from as soon as they can breathe on their own (Oller, Caskey, et al., 2019). Human infants produce protophones at least ten times more

frequently than chimpanzees and bonobos produce sounds viewed as analogous to protophones (Oller, Griebel, et al., 2019). Spoken words begin at the end of the first year, but human infants use protophones to communicate needs and states of being from the first month (Gratier et al., 2015; Jhang & Oller, 2017), long before gestural communication has been reported. The importance of vocal communication is emphasized by evidence that gaze-coordinated vocalizations are stronger predictors of later language outcomes than either gestures alone or gesture-vocal combinations (Donnellan et al., 2020).

Long et al. (2020) emphasized the endogenous nature of infant vocalization; in the first year infants produced three times as many protophones independently as during social engagement. From 9 to 18 months, babbling practice alone, unaffected by social environment, has been found to be a strong determining factor for word onset (McGillion et al., 2017).

An Evolutionary-Developmental Perspective

Evolutionary-developmental (evo-devo) biology emphasizes the widespread tendency for new structural features or capabilities to evolve by modification of developmental patterns (Müller & Newman, 2003). In evo-devo theory, conservation of foundational structures is expected, and natural selection is seen to build upon the foundational structures (West-Eberhard, 2003). Thus, the order of appearance of structures or capabilities/activities in development is expected to emerge following evolutionary orders (Carroll, 2005; Newman, 2016). In accord with this line of thought, one should predict that if gesture forms the primary foundation of language, then gestural communication should predominate in early communication in humans, and conversely if vocalization forms the primary foundation, then vocalization should predominate.

Aims

Although there exist well-established procedures for judging both the structure and communicativeness of protophones from birth and across the whole first year, we know of no descriptive framework identifying gestures and their potential communicative roles that is applicable to the whole first year. In the present work, we propose a framework to allow comparable counting of communicative and/or potentially communicative events of both infant gesture and vocalization. Thus, we aim to provide new perspectives on the relative roles of gesture and vocalization in modern human communicative development and indirectly in the evolutionary origin of language. We expect that empirical data will contradict the expectation, based on the gestural origins theory, that gesture should occur more frequently than vocalization in the first year. We hypothesize the opposite:

1) speech-like vocal events (protophones) will occur at a higher rate than gestures in early human development and
2) protophones, more often than gestures, will show signs of constituting intentional communications, being accompanied more often by gaze directed toward another person.

Methods

Selection of Participants

Data were acquired from archived longitudinal audio-video recordings from the University of Memphis Origin of Language Laboratory. Approval for the research was obtained from the University of Memphis Institutional Review Board for the Protection of Human Subjects (IRB). All participants resided in or around Memphis, Tennessee. Recruitment was conducted in child-birth education classes and by word of mouth. All infants' parents completed a written consent approved by the IRB prior to recordings. An inclusion criterion for

participation was normal pregnancy. Typical development (i.e., lack of hearing, vision, language, or other developmental disorders) was confirmed throughout participation in the study via parent report using information such as passed hearing screenings and mastery of developmental milestones at expected ages. None of the infants was born prematurely.

The archived recording sessions included 21 parent-infant dyads from two waves of longitudinal study (see e.g., Oller et al., 2013; Oller, Caskey, et al., 2019). Based on funding available for coding and analysis, we were able to select only a subset for the present research, with recordings from 10 infants (5 male, 5 female) balanced for age, gender, recording session type, and recording length. The recordings from which the data were drawn usually included three sessions, each approximately 20-minutes in length, often based on a single continuous ~60-minute recording. We analyzed only the sessions termed "interactive" at approximately 4, 7, and 11 months—thus we coded 30 sessions, one for each infant at each age.

In the interactive sessions, parents were instructed to engage their infants as they normally would, whereas the other two recording sessions during the ~60 minutes required the infant to be present with the parent reading (the no-talk-to-baby circumstance) or playing separately while the parent was interviewed by another adult. We selected the interactive sessions for our study, assuming gestures would more likely occur in those sessions since parents were more likely to look at infants during those sessions than during the no-talk-to-baby or interview sessions. Thus, selection of the interactive sessions served our intention to maximize the occurrence of gestures.

We selected infants with recordings fitting the age criteria to the extent possible while taking advantage of existing vocalization and gaze coding from prior work. All but one of the infants were White (Infant 4). All were learning English as their native language except for

Infant 2, who was exposed to German, Spanish, and English. All infants were of middle to low-middle socio-economic status. Demographics and recording ages are provided in Table 1. Selections were made without regard to rate of occurrence of gestural or vocal activity.

Table 1. Infant Demographics and Ages at Recordings
This table displays demographics and recording ages in months and weeks for each infant at each session.

Infant	Gender	Race	Age at Recordings (months; weeks)		
			Early	Middle	Late
1	F	White	3;0	5;0	10;1
2	F	White	4;2	6;0	11;2
3	M	White	5;1	7;2	11;1
4	F	Black	3;1	7;1	12;0
5	M	White	3;3	6;3	10;1
6	F	White	3;2	7;1	10;2
7	M	White	3;3	7;1	11;3
8	F	White	3;3	7;0	12;2
9	M	White	3;3	7;1	11;3
10	M	White	3;3	7;0	11;3
Average age in months; weeks		*M* (*SD*)	4;0 (0;3)	7;1 (0;3)	11;2 (0;3)

Data Collection

One recording was selected for each infant (total: 10) at each age (total: 3), yielding 30 recordings. The actual length of the planned 20-minute sessions was ~19 minutes (range: 16 - 20 minutes). All parent-infant dyads were recorded in a quiet laboratory/playroom with toys and books appropriate for the infant age. Both infants and parents wore high fidelity wireless microphones. The playroom was equipped with cameras in the corners, either four cameras or eight (one high camera and one low camera in each corner). The differing number of cameras corresponded to three phases of recording in three different laboratories. Laboratory staff operated the cameras for zoom and tilt from an adjacent control room. Two of the four or eight channels were selected to record the interaction at each moment. Selection and zooming afforded

close-up views of the infant face, torso, and actions on one of the selected channels and a broader view of the interaction (including the parent) on the other channel. Details regarding laboratory equipment and procedures can be found in previous work from this laboratory (Buder et al., 2008; Oller et al., 2013). Instructions to parents for the interactive segments emphasized playing with and interacting with infants in a natural way, allowing for vocal, gestural, and tactile interaction at any time (additional information on recordings in SM, 2.1).

Coding Approach and Rationale

Coding Rationale

Gesture and vocalization are not easy to compare given that the two modalities have different advantages and disadvantages. Still, it is both appropriate and necessary to quantitatively compare gesture vs. vocalization across development in order to address the relative roles of gesture and vocalization in language origins (see SM, 1.1). We have striven to construct an approach allowing well-motivated quantitative comparison. We sought to ensure that comparisons would not bias outcomes in favor of vocal acts; on the contrary, we sought to ensure that any bias would favor gestures (see SM, 2.3). We selected an event-based analysis, not a duration based-analysis. This choice is motivated by the fact that *events* of communication or potential communication are the optimal points of comparison, since each gestural or vocal event is a potential communicative act (for duration comparisons see SM, 3.2).

Software Environment

Coding was conducted in AACT (Action Analysis Coding and Training software, Delgado et al., 2010), a software environment facilitating simultaneous coding of video/audio. AACT presents two channels of video synchronized with audio at frame-level accuracy, with audio displayed spectrographically and in waveform through a version of TF32 (Milenkovic,

2010) designed for AACT. The coding fields of interest for the present research were protophone types, gestural act types, gestural illocutionary functions, and gaze direction accompanying both protophones and gestures (see SM, 1.3, 2.4 and 3.4). Regarding facial affect coding, see SM, 3.3. Aside from gesture coding, data collection was completed in a way that was identical to previous infant vocalization studies conducted in the Origin of Language Laboratory (Jhang & Oller 2017; Long et al., 2020; Oller et al., 2013). During coding, the coder places a start and an end cursor on the TF32 spectrographic display and can play the selected sequence repeatedly in an AACT "loop". The cursor can be dragged on the spectrographic display or shifted with keyboard controls producing frame accurate shifting in the video from both cameras to enable selection of onset and offset points for events in any field.

Vocalization Coding

Vocalizations had been coded in prior research (e.g., Oller et al. 2013, Oller, Caskey, et al., 2019). The primary focus of this previous work was speech-like vocalizations, focusing on phonatory properties. This approach resulted in three primary types: vowel-like (vocant), growl-like, or squeal-like sounds. Oller (2000) refers to these types as "protophones" (including both canonical and non-canonical sounds). Thus, syllables or syllable sequences such as "dada" or real words were, like precanonical protophones, categorized in terms of phonation as vocant, growl or squeal. Cries, laughs, and whimpers as well as a variety of additional infrequently occurring types (whispers, voiceless friction sounds, ingressive sounds) were also coded but not included for analysis.

Coding was conducted with repeat-observation, assigning boundaries in TF32 at the onset and offset of each protophone. Cursor placements for each utterance were recorded in AACT, specifying duration in ms. Boundaries were assigned using a "breath-group" criterion (Lynch et

al., 1995); thus an utterance was determined to begin with phonation during exhalation (i.e., egress) and end with termination of phonation, often accompanied by inhalation (i.e., ingress). Thus, a new utterance could begin only after an observed breathing pause. A protophone type was then selected from the coding panel, and in a subsequent coding pass, a gaze direction category was assigned to the time period of each vocalization. If any gaze was directed to a person during the utterance, even if only briefly, the utterance was coded as having been directed to a person. Details on vocalization and gaze direction coding and associated coder agreement can be found in prior publications, especially in their supplementary materials (e.g., Jhang et al., 2017; Oller et al. 2013).

Gesture Coding

Global Categories for Gesture Coding and their Relation to Global Vocalization Categories

The coding scheme was intended to maximize comparability between events in the gestural and vocal domains. We sought to categorize gestural acts during the first year that could be considered precursors to signs (i.e., precursors to gestural symbols) just as we categorized vocal acts that could be considered precursors to words. The coding consisted of four global movement categories: 1) Utilitarian Acts, 2) Non-social Gestures, 3) Universal Social Gestures, and 4) Conventional Gestures. We defined *Utilitarian* Acts as those actions that are simply world-exploratory or manipulative, without any inherently social communicative function. For example, if a child reached for and obtained any object, the event was coded as a Utilitarian Act. These acts were not counted as gestures even though they were coded in the initial real-time coding pass. Vocalizations also include Utilitarian Acts in the form of vegetative sounds (coughing, burping, clearing one's throat, blowing out a candle…) that perform functions related

to respiration and digestion—vegetative sounds were not counted as protophones.

The remaining three global gestural categories were intended to include all the acts that could conceivably be interpreted as communicative or those that could be expected to be brought into the service of communication at some later point in development. *Non-social Gestures* are gestural actions that are not merely Utilitarian and are also not inherently communicative, although they have *the potential for being utilized communicatively*. For instance, rhythmic hand banging and foot tapping are examples of Non-social Gestures, akin to babbling/protophones in the vocal domain, actions that can eventually be brought to the service of communication, but that are not yet intentionally communicative. Non-social Vocalizations included all the protophones—the very infrequently occurring words were categorized as both words and protophones, but words were treated as Social, while non-word protophones were treated as Non-social.

Universal Social Gestures include acts with an inherently social communicative intent but with no reason to presume they are learned from specific cultural experience. For example, an extended flat hand to indicate refusal is a Universal Social Gesture. Also, reaching upward with both hands when an infant wishes to be picked up is a Universal Social Gesture. Pointing and other clearly deictic gestural acts are also Universal Social Gestures.

Universal Social Gestures have fixed functions in infancy; e.g., deictic gestures designate entities but cannot name them or perform other affectively-valenced functions such as expressing distress or delight (see SM, 1.2). Other Universal Social Gestures, including reaching toward someone to request being picked up or a flat hand outstretched to indicate refusal are also fixed in function, in that they cannot acquire a different function through learning in infancy. Similarly, there are *vocal* acts in infancy that have fixed functions (the Universal Social

Vocalizations), crying and laughter, which express negative or positive affect respectively, but cannot designate objects, as deictic gestures can. While we counted Universal Social Gestures in the quantitative comparison of gestures and protophones, Universal Social Vocalizations were not included (see SM, 2.3).

Conventional Gestures are those that are culturally transmitted, acts with a discernible communicative function, such as waving to convey "hello" or "bye-bye", clapping in celebration, or thumbs up to indicate approval or agreement. These gestures are learned and can be viewed as analogous to primitive words, i.e., Conventional Vocalizations. Non-word protophones (including both non-canonical and canonical babbling) can, in accord with our scheme, be subcategorized as Non-social, while words (treated here as a subclass of protophones) are subcategorized as Conventional Vocalizations (speech). The four global categories, then, facilitate comparison across vocalization and gesture since both the gestural and vocalization coding schemes presume the same four.

It is important to recognize comparable aspects of the two modalities while also taking account of inherent differences in how the two modalities can function (see SM, 2.2). Universal Social acts, both of gesture and vocalization are communicative but require no associative learning, and their intended functions are interpretable to potentially everyone around the world. Universal Social Gestures are capable of transmitting certain critically important communicative functions such as refusal, request, and designation (pointing), while in the exclusively vocal domain, these functions seem impossible to transmit without symbols (words), and even then, vocal transmission is more complicated. But vocalization has the advantage of being able to transmit affective valence (Jhang et al., 2017), which is difficult if not impossible to transmit by gesture alone. Thus, a fundamental difference between the modalities is that vocalization is well-

suited to transmission of emotional valence, while gesture is well-suited to transmission of deixis.

Consider designation, which is possible with words (“look to your left and notice an orange object”), but not with protophones. Even in young infants, a simple act of pointing can serve as a designation (directing attention to the orange object without a single word). Similarly, other functions that can be transmitted with gesture universally even in infancy (refusal or request, for example) cannot be uniquely transmitted in vocalization without words (“I don’t want it”, “stop”, “give me the book”). Still, prosodic features of vocalization and/or facial affect can emphasize or modulate the flavoring of such functions in either modality and regardless of whether the communication is prelinguistic or based on signed or vocal symbols.

There is thus a gap in the potential for communication transmitted with protophones, a missing deictic function, a gap that can be at least partially filled by Universal Social Gestures, which consequently provide a scaffold for early communication and especially for learning of labels. This special capability of Universal Social Gestures may account for the widespread opinion that human language is founded in gesture. Yet, vocal communication begins in the first week of life, while gestural deixis does not appear until late in the second half year. The gestural-origin opinion also neglects the fact that gesture has a gap in transmission of affective valence, a gap which is partly filled with Universal Vocal Acts such as crying, whimpering, laughter, and whining, which may also supply support for the emergence of language. Furthermore, deictic gestures, while supplying a support system for learning symbols in either the gestural or vocal domain, are not themselves symbols (SM, 1.2).

Gesture Coding Procedures

The primary coding was conducted by the first author. We drew a distinction between gestural *actions* and the potential *intent*, or communicative function, associated with each action. This distinction was also drawn between *vocal* actions and their communicative functions. For example, "reaching" is a code for an action, but that code does not characterize the communicative function of the action. An infant could reach to show an object, to request an object, to try to get the attention of another person, or to offer or accept something. Our coding scheme included two fields, in one case for each individual gestural action and in the other for the communicative function of each action.

The actions and functions were coded for each ~19-minute segment in three separate passes using both audio and video. The first pass, using real-time observation, allowed the coder to acquire an overview of the recorded segment and to mark the approximate location of infant gestures, labeled by keystrokes in real-time, each corresponding to one of the global codes: Utilitarian, Non-social, Universal Social, or Conventional. In the second pass, the coder ignored Utilitarian Acts but used repeat-observation to designate boundaries for individual Non-social, Universal Social, or Conventional gestural actions at the onset and offset of each action, specifically designating the point of the movement beginning the event through the point where the event reached its full extension. For example, the onset of an index-finger point begins when the arm, hand, or finger moves into motion from rest until the moment where it reaches its full extension. In cases of rhythmic movements such as hand banging, boundaries were determined using a criterion similar to the breath-group criterion for vocalizations. Thus, duration of any cluster of actions (such as the individual strokes of rhythmic hand banging), deemed to constitute

a gesture occurring before a gestural pause (treated similarly to pauses in speech), specified the duration of the gesture.

In the second pass of gesture coding, the cursors in TF32 were adjusted to make each boundary decision, which sometimes required the cursors first to be set as much as 1000 ms before the onset and after the offset of the gesture to allow extended viewing of the event surroundings. Then the cursors could be moved (by dragging them in the acoustic display with corresponding frame accurate changes in the video or by keystrokes that could move the video display one frame at a time with a corresponding shift in the audio display) to home in on the actual boundaries of the gesture specifying the onset and offset points. A keystroke indicating a particular gestural act was then recorded in AACT, indicating both onset and offset times.

Once onset and offset boundaries had been determined, we used the bounded time frames of the gestural actions to automatically create placeholders in a new coding panel (the gestural function panel). The placeholders were recorded sequentially in the gestural function panel (without showing the gestural action codes) and allowed the coder to categorize all the bounded events in a third pass, where a gestural function was designated for the precise time frame of each gestural action, the entire period of each action being taken into account in coding the function. The second and third coding passes for actions and functions used the detailed gestural category labels as listed in Table 2 and Table 3, respectively.

Table 2. Gestural Action Categories Used in the Present Study

The list is intended to include actions that could conceivably be interpreted as communicative gestures. The list is not complete, but includes all the gestural types actually observed. The terms are drawn partly from literature on ape and human infant gesture.

Global Category	Gestural Action	Definition
Non-Social	Hand Shake	Shaking or flapping of hands with no explicit social intent
	Hand Position	Posturing of hand intentionally, such as creating a "d" or an "f" hand (as in American Sign Language)
	Body Rock	Rocking body back-and-forth or bouncing in an upward and downward motion
	Foot Shake	Shaking or flapping of feet with no explicit social intent
Universal Social	Reach	Extension of arm away from body, stretching toward an object, person, or surface
	Point	Extension of index finger while remaining fingers flex into the palm ("d" hand)
	Arm up	Extension of arms and hands in an upward motion toward another person
	Throw	Throwing object to someone
	Push	Pushing object to someone
	Block	Extension of hand or arm to prevent any contact with one's person by another individual or object
	Touch	Extension of hand and/or arm to gently make contact with another person
Conventional	Clap	Striking palms together repeatedly
	Cover Face	Hand(s) or object (e.g., a blanket or cloth) placed over face to obstruct another's view of face
	Hand Wave	Movement of hand(s) or entire arm back-and-forth with palms facing away from body
	Head Shake	Shaking head from side to side in a continuous motion

Table 3. Gestural Function Categories

All Non-social actions from Table 2 were treated as expressing Non-communicative functions. Universal Social and Conventional actions were categorized as expressing one of the Communicative functions. Depending on the apparent intent of the infant, Reach could be categorized as expressing Request Object, Offer, or Accept. As in Table 2, the Communicative function list is not complete, but includes all the gestural functions actually observed.

Function Category	Gestural Function	Definition
Non-communicative	Non-social act	Any non-utilitarian act, not conveying communicative intent (e.g., rhythmic hand banging, body rocking)
Communicative	Request object	Show desire to obtain something
	Accept	Receive something offered
	Social playful	Engage in interactive game, usually involving an object (e.g., playing catch)
	Offer	Present something to someone to accept/reject
	Request up	Show desire to be held or picked up
	Designate	Indicate person or object of interest
	Exult	Show happiness or excitement, celebrate
	Bye-bye	Wave "bye-bye"
	Refuse	Indicate unwillingness to do something
	Show	Make something visible to be perceived by another
	Seek attention	Show desire to engage with another
	Assent	Express approval or agreement

Gaze Coding

Rationale for Gaze Coding

Previous studies have argued that gaze coordination can be used as an indicator of intentionality in prelinguistic vocalizations and gestures (Bates, 1976; Donnellan et al., 2020; Harding & Golinkoff, 1979; Iverson, 2010; Iverson et al., 2000; Thal & Tobias, 1992; Wu & Gros-Louis, 2015). Gaze direction is an excellent predictor of judgments of social directivity based on the conjunct of factors (timing, prior social context, etc.) that suggest an infant as trying to communicate (Long et al. 2020).

Gaze Coding Procedure

In a fourth pass, gaze direction was coded during each vocalization and during each gesture where at least one of the two video views allowed such judgement. The coding determined whether infants produced a protophone or a gesture while looking at another person (i.e., a socially-directed communication) or while not looking at another person (i.e., a non-socially-directed communication).

As indicated above, the gaze direction for protophones had been previously coded during other studies from our laboratory, and those judgments were used in the current study. Following the same coding procedure as in the third pass of gesture coding, coders clicked on each individual placeholder event indicating a previously coded gesture or vocalization and then coded the direction of the infant's gaze during the bounded action plus 50 ms on either side. Playback thus started automatically 50 ms before the event onset and continued through 50 ms after the offset. This procedure was designed to ensure that all video frames of each event would be viewed before determining the directedness of the infants' gaze. The categories for coding were 1) directed toward another person, 2) not directed toward another person, or 3) can't see

(sometimes the infant's gaze direction could not be judged based on either camera view). If there was any moment of gaze direction to a person observed during the gesture or vocalization interval (the whole event plus the 50 ms additions), the event was coded as person directed.

Coder Agreement Training and Agreement Outcomes

Two graduate students were trained as agreement coders for gestural actions, gestural functions, and gaze direction. The training began with a lecture by the first author on the gesture coding scheme, during which coders were presented with examples of video-recorded infant gestural actions previously coded by the first author and confirmed by the last author, all of which either met a consensus standard for one of the gesture categories or displayed ambiguities of possible judgements deemed instructive for training. Gaze direction training was similar. Once training was completed, coders followed the same coding procedure as the first author, using the criteria outlined in the gestural coding scheme and the gaze direction scheme. The first author selected 12 recordings semi-randomly from the 30 recordings for the agreement study, with one five-minute segment selected from within each of the 12 recordings. Four samples came from each of the three ages and all ten infants were represented in the agreement samples. Neither of the agreement coders knew the hypotheses for the study and were blinded to coding of the first author.

There was high agreement for the number of gestures identified in the five-minute segments for both agreement coders with respect to the primary coder (r = 0.92, r = 0.81, N = 12), and for the two with respect to each other (r = 0.80). On proportion of gestures with gaze directed to a person, the agreement coders also showed good agreement with respect to the primary coder (r = 0.92, r = 0.73) and each other (r = 0.72).

Across the five-minute segments in the agreement samples, all three coders showed at least a doubling of the rate of gesture from the first to the second age, with an increase ranging from ~50% to 60%. The three coders also showed an increase ranging from 22% to 65% in the amount of gesture from the first to the third age. Also, the three coders showed very similar proportions (27%, 30% and 30%) of person-directed (by gaze) gestural actions. For all three coders, 13% to 14% of gestures were deemed to be directed to a person, whereas degree of directivity for protophones was > 27% for all three. The agreement data suggest that if the entire data set had been coded by either of the agreement coders instead of the primary coder, none of the conclusions associated with the results reported below would have changed.

Results

The Hypotheses

Hypothesis 1, Distribution of Gestures and Protophones

The data on protophone and gesture rates (Figure 1) across all 30 recordings revealed vastly more protophones (3903) than gestures (752), with gestures occurring infrequently (only about one every 4 minutes) at the earliest age, increasing to 1.7 – 2 per minute at the later ages. In contrast, protophones occurred approximately 10 times per minute at the youngest age, and nearly 6 per minute at the oldest. Thus, there were more protophones than gestures, and this was especially true at the youngest age. These results support a more vocal than gestural origin for communication.

Shapiro Wilks tests for normality on the gesture and protophone rates were run separately. Based on the tests, protophone rates and protophone proportions of social directivity can be assumed to be normally distributed, but gesture rates and proportions of social directivity

were far from normal, primarily because of zeros at the earliest age. Consequently, a non-parametric approach to repeated-measures analysis was necessary.

Data were analyzed with Generalized Estimating Equations, a nonparametric alternative to generalized linear mixed models, producing unbiased regression estimates for use in longitudinal or repeated-measures research designs with non-normal response variables (Liang & Zeger, 1986; Ballinger, 2004). Indeed, the data in question were nonnormal especially because of the very low and unequal numbers of gestures at the early Age. We determined GEE was appropriate also because of the unequal amounts of data in the two Modalities across Age and lack of precise Age matching across infants. The model included two Modalities (gestures vs. protophones) and a factor with three Ages (allowing comparison of early to middle and early to late), with 10 infants at each Age. The dependent variable was number of events produced, either gestures or protophones. The covariance matrix was exchangeable. With an unstructured matrix, the GEE failed to converge. The robust sandwich estimator was used to obtain standard errors of estimates.

GEE revealed a significant early to middle Age by Modality interaction ($p < .02$) as well as a stronger early to late Age by Modality interaction ($p < .0005$), reflecting the dramatic increase in the number of gestures across Age, in contrast to protophones, which showed no such increase, but in fact fell in frequency across Age, though not so dramatically as gestures rose. A significant main effect of Modality ($p < .00001$) reflected the much larger number of protophones than gestures. Main effects of early vs. middle Age ($p < .00001$), and early vs. late Age ($p < .00001$) reflected a mean increase in the combination of gesture and protophone rates across age.

Because of lack of uniformity of ages within the groups (see Table 1), we also ran GEE with Age as a continuous variable. The results indicated a significant Age by Modality interaction ($p < .005$), reflecting the fact that gestures increased dramatically with Age, while protophones fell. There was also a significant main effect for Modality ($p < .00001$), as in the analysis with Age as a three-level factor, and a significant effect for Age ($p < .00001$), again reflecting growth in the combination of gesture and vocalization across Age. GEE details are in SM, 3.1, Table SM1 and SM2.

Follow-up Mann Whitney U tests (non-parametric) also showed a significant difference reflecting more protophones than gestures (z score = 3.74, $p < .0002$). In addition, Mann Whitney U tests confirmed that gesture rate was lower at the early than at the middle (z score = 3.10, $p < .002$) or late Ages (z score = 3.36, $p < .001$). The difference between the middle and late Ages was not significant. For protophones, rates fell across Age, but only the comparison between early and late Ages was significant (z score = 1.97, $p < .05$).

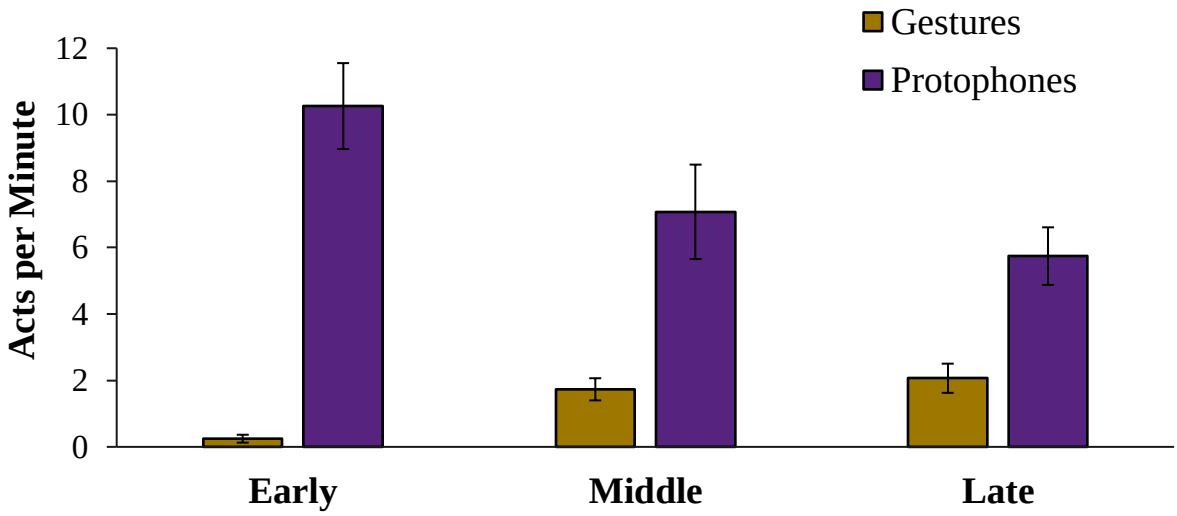

Figure 1. **Gestures and Protophones per Minute Across the First Year**
This figure shows protophones and gestures per minute across the three ages in our sample (error bars show standard errors).

Hypothesis 2, Directivity of Gestures and Protophones

The data on directivity of gestures and protophones indicated, consistent with Hypothesis 2's prediction, that protophones were more likely than gestures to be socially directed as indicated by gaze direction toward a person. Averaged across ages, the grand total of protophones was nearly twice as likely to include person-directed gaze as gestures (34.6% to 17.7%). The tendency was greatest at the youngest age (37.6% to 21.7%) and least at the latest age (30.2% to 17.7%). Figure 3 presents averages at the infant level, illustrating a tendency for the directivity proportion to be highest at the latest age for gestures, but lowest at the latest age for protophones.

To compare proportion of directed protophones vs. gestures statistically, we selected a similar GEE model to the one used for Hypothesis 1, with Age as a three-level factor. The dependent variable was proportion of events in each Modality directed by gaze to a person (Figure 3). GEE revealed a significant interaction of early to late Age with Modality ($p < .05$),

reflecting conflicting tendencies; gestures showed greater directivity at the late than the early Age while protophones showed the opposite. There was no significant difference for the interaction of early to middle Age with Modality. As with the test for gesture and protophone counts, GEE revealed a very strong main effect for Modality ($p < .00001$), indicating that protophones were more often socially directed than gestures. The result again supports the prediction that protophones are more socially-directed than gestures, diametrically contradicting the gestural origins theory prediction that early gestures should be more socially directed than vocalizations. The analysis also revealed a main effect for early vs late Age ($p < .05$), reflecting an overall tendency for directivity to be higher at the late Age for gestures and protophones combined.

As with the analysis of gesture and protophone counts, we conducted a separate GEE analysis of directivity proportions with Age as a continuous variable rather than as a three-level factor. The results indicated a significant interaction of Age by Modality ($p < .00001$), a significant main effect of Modality ($p < .00001$), and a weaker significant effect of Age ($p < .02$). Details of these GEE analyses are in the SM, 3.1, Tables SM3 and SM4.

Follow-up Mann Whitney U tests also showed the proportion of socially-directed protophones was greater than that of gestures (z score = 2.91, $p < .001$). The same pattern occurred for early (z score = 2.78, $p < .006$) and middle Ages (z score = 2.38, $p < .02$) but was not significant at the late Age. Follow-up tests also showed gesture had higher directivity at the late than the early Age (z score = 2.15, $p < .05$), but other Age comparisons were not significant. No Age comparison by Mann Whitney U test was significant for proportion of protophone directivity. The great majority of all events were not directed toward another person in *either* Modality.

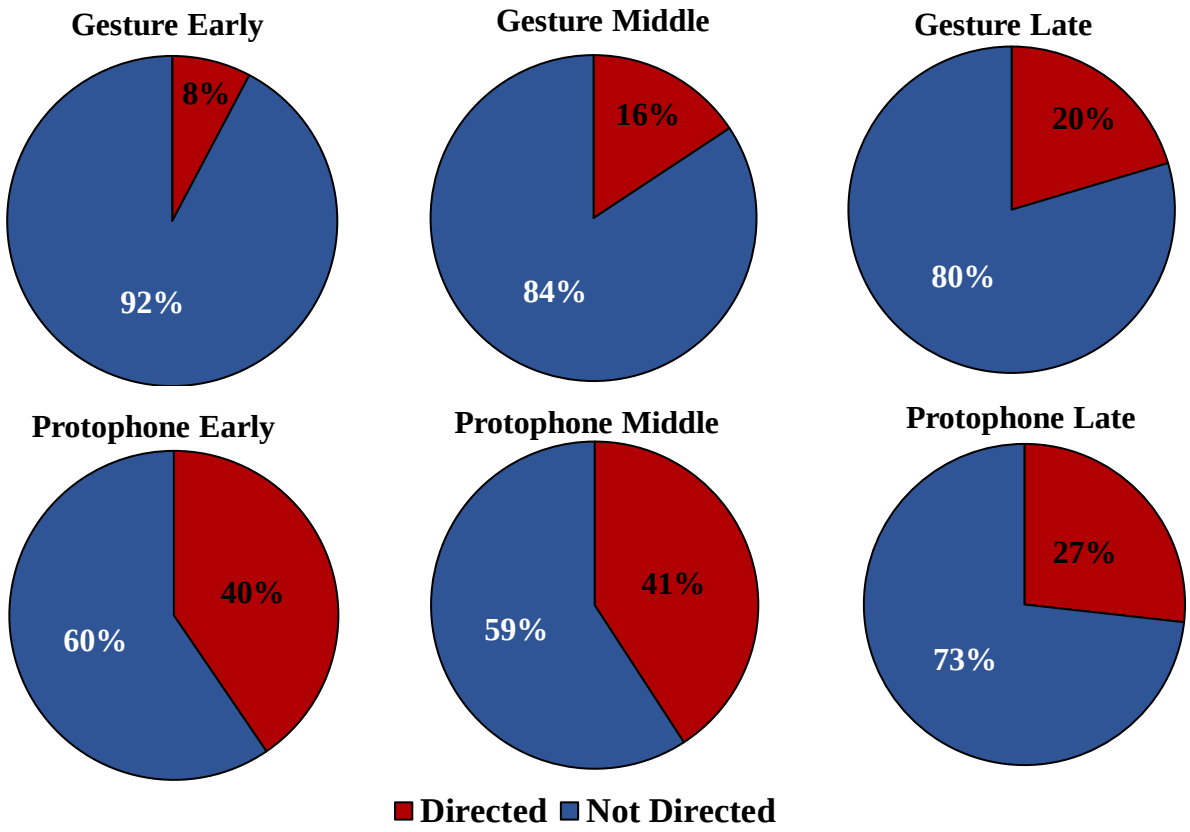

Figure 2. Proportion of Directed Protophones and Gestures
This figure shows proportions of protophones and gestures that were socially directed as determined by gaze direction at three ages.

Distribution of Gesture and Protophone Types

Table 4 shows the distribution of three global gesture categories. Numbers of Non-Social and Universal Social Gestures at the early and middle ages were similar and were by far the most frequent gesture types at the early and middle ages. At the late age, ~60% of gestures were Universal Social Gestures. Pointing only occurred 12 times in the ~600 minutes of coded recording, 9 times from one infant at 11 months, and three times from another infant. Both at the middle and late ages, the great bulk of Universal Social Gestures involved reaching as if to

request an object, offering an object to another person by reaching, or accepting an object by reaching.

Conventional Gestures were not frequent at any age (42 total gestural events, 26 of which occurred at 11 months). Those occurring at 11 months included 4 types: no-no (head or finger shaking), hand clapping, bye-bye (waving), and face covering. All Conventional Gestures at the middle age were associated with a peekaboo game with one infant only.

Although the available coding of protophones did not specify them for the three global categories (social and non-social usage of vocalizations had not been coded in the earlier studies), we evaluated tokens of Conventional Vocalizations across the three ages. We found 50 conventional words (including mama, dada, bye-bye, no-no, yum-yum, mmm (tastes good), and yeah), compared with the 42 Conventional Gestures.

Table 4. Distribution of Gesture Types
This table shows the distribution of three global gesture categories with total gestures across the ~600 minutes of recording.

Gesture Type	**Infant Age Group**			
	Early	Middle	Late	Total
Non-Social	25	164	142	**331**
Universal Social	21	149	206	**376**
Conventional	0	16	26	**42**
Total	**46**	**329**	**374**	**752***

Discussion

Outcome Summary

The present study provides the first direct comparison of gesture and protophone rates across the first year. The infants produced more than five times as many protophones (3903) as

gestures (752), with a substantial imbalance favoring protophones at all three ages. The vocal predominance applied whether the communication was directed or not directed by gaze toward another person. There were 1,349 cases where a protophone was directed to a person by gaze, compared with 133 cases of directed gaze for gesture. Furthermore, protophones were nearly twice as likely to be person-directed as gestures (34.6% to 17.7%). If then, we take gaze direction as an indicator of communicative intent, we can conclude that not only did protophones occur much more frequently than gestures, but protophones were far more likely to be intentionally communicative than gestures. The results diametrically contradict the gestural origins theory: early communicative development was not dominated by gestures, but by protophones.

The interactions between Age and Modality indicate that gesture not only became more frequent across the first year, but also became more socially directed, while the opposite occurred for protophones. We are doubtful about the replicability of the apparent fall in protophone rate across Age, partly because the effect was only statistically significant from the early to the late Age, and also because other research has not clearly found such a pattern (Gilkerson et al., 2017; Oller, Caskey, et al., 2019; Iyer & Oller, 2008; Iyer et al., 2016). But the significant interaction suggesting a rise in directivity of the gestures and a seeming fall in directivity of the protophones (patterns that have to our knowledge never before been addressed) inspires us to consider an interpretation based on the possibility that infants learn about the communicativeness of their gestures and protophones across time and that they adjust their gaze direction accordingly. Namely, the results suggest the possibility that infants learn across the first year that someone needs to be looking at them for their gestures to be communicatively effective,

while also learning across the same period that their vocalizations can often be communicatively effective even without anyone looking at them.

Inherent Differences Between the Gestural and Vocal Modalities

At the late age, most gestures were Universal Social Gestures, highlighting a particular communicative feature of the gestural modality that is difficult to implement vocally (see SM, 1.2). Universal Social Gestures communicate intents, e.g., indicating something the infant wants or may want another person to look at. In contrast, vocalizations must first become symbolic in order to serve such functions. An infant can gesture by extending his or her arm(s) to signal the desire to be picked up, but there is no equivalent in the vocal domain without words (e.g., "pick me up").

On the other hand, vocalization can transmit emotional valence, and thus can be used universally to modulate the affective tone of communication. It is universal for caregivers to recognize cry and fussy protophones; each protophone type can be flavored by intonation or other acoustic modulations to convey affect. Approximately 15% of protophones in laboratory recordings labeled as fussy can be judged as negative from sound alone (Jhang & Oller, 2017). Thus, protophone prosody can assist in flavoring affect of communication in a way gestures cannot, but prosody cannot supplant a gesture's deictic function. Facial affect can be utilized to modulate the emotional tone of communication in either gesture or vocalization.

Vocalization can be used to assist in gestural communication by supporting attention seeking (Franco et al., 2008; Gros-Louis & Wu, 2012). Pointing or reaching are often accompanied by vocalizations to attract listener attention. But even in these cases, the vocalizations (unless they are words) cannot serve the deictic functions that are natural to the gestural domain.

An important feature of protophones is functional flexibility (Oller et al. 2013). This feature contrasts sharply with the functions associated with Universal Social Vocalizations or Universal Social Gestures. Crying, for example, is naturally associated with a function of distress expression. Pointing is naturally associated with a designative function. All universal communicative acts have a specifiable function or class of functions. In contrast, no protophone type has a universal function—all protophones can be produced to serve multiple functions. It is a critical feature of protophones that they must be free to serve functions with all possible valences, because if they were not free in that way, they could not form a foundation for symbolic words, which are by definition free of any particular illocutionary function.

Of course, both gesture and vocalization are capable of developing into full-fledged language with full functional flexibility. Both modalities include actions that are not inherently tied to particular universal functions. With both Non-Social Gestures and protophones, infants explore actions free of particular function; only with this freedom can they be adapted at a later point by learning to form Conventional acts.

The flexibility of protophones is emphasized by recent results from coding of laboratory recordings, where it was found that ~75% of infant protophones are produced without social directivity (Long et al., 2020). This fact suggests protophone production is largely endogenous. It has been reasoned that protophone production offers caregivers information about an infant's well-being (Locke, 2006 and see SM, 1.3). The same kind of reasoning may be thought to apply to gestural babbling, i.e., to Non-Social Gestures, but if gesture formed the primary foundation for language, we would not expect infants to look towards caregivers only about half as often during gestures as during protophones.

Interpretations of Unanticipated Findings

The great majority of gestures observed were not the ones expected based on the common suggestion that language is founded in gesture. Pointing, known to form a foundation for word learning, occurred far less frequently than expected. Only 12 pointing events were observed (<2% of observed gestures). An even lower frequency of pointing was found in recent work on communicative vocalizations and gestures at the end of the first year in a study of 134 prelinguistic infants (Donnellan et al. 2020), where only 28 cases of indexical pointing were observed, and many of those were not gaze-coordinated.

In our own data, reaching accounted for 48% of all gestures, suggesting that the "declarative" function of pointing, thought to be so important as a foundation for language (Bates et al. 1979), occurs far less frequently than the instrumental functions associated with reaching through 11 months of age. Only one other category of gesture occurred frequently. Rhythmic hand shaking, which we interpret as a Non-social Gestural act similar to reduplicated babbling, accounted for 35% of gestures overall and 26% at 11 months. Together, hand shaking and reaching accounted for 83% of gestures.

Both Conventional Gestures and Conventional Vocalizations were infrequent in the first year, with slightly more Conventional Vocalizations (words) than Conventional Gestures, and both the Conventional Gestures and the words were overwhelmingly performatives, that is, they constituted illocutionary acts such as greeting (waving, saying hello), celebrating (clapping, saying hooray), or refusing (head shaking, or saying no), rather than semantic acts of reference, such as naming an object or describing an event. Thus, the Conventional acts in both modalities were overwhelmingly dyadic, constituting communications between two parties with respect to

each other, rather than being triadic communications where the two parties jointly referred to a third entity through a conventional symbolic act.

Evaluating Events of Communication, Individual Differences, and Evolutionary Implications

Although the average duration of a protophone was only about half as great as that of a gesture, protophones in the recordings accounted for considerably more time than gestures (SM 3.2). Comparison at the event-level is more useful than duration comparison, because each event in either modality constitutes a possible communicative act, and it is the number of such possible communications that matters most in assessing the relative importance of gesture and vocalization.

There was notable variation among the 10 infants in protophone rate, perhaps due largely to natural day-to-day variability, but the variation is intriguing nonetheless. Even in the vocal domain, individual variation was salient, with a low of 35 protophones at the middle age from one infant to a high of 281 from a different infant at the early age. Of special interest was the fact that only two infants accounted for >70% of all gestures at the early age, and those two were the only infants who were >3 months at the time. The largest number of gestures among the other 8 infants at the early age was 4, and five infants produced either 0 or 1 gesture at that age. Consequently, it is tempting to speculate that human infants do not significantly engage in gesture until after 3 months.

In spite of substantial individual variation, the key differences were robust. All 10 infants produced more protophones than gestures; the infant producing the fewest protophones produced 1.9 times more protophones than the gestures produced by the infant producing the most gestures. 9 of the 10 infants produced a higher proportion of person-directed protophones than

person-directed gestures; even the infant who produced a higher proportion of person-directed gestures than protophones overall, did so only at the late age, while all the other infants showed a higher proportion of person-directed protophones than gestures at all three ages.

Gestures and vocalizations did not tend to co-occur. Only 17% of gestures overlapped with a protophone (see SM, 3.5). Thus, the data did not suggest extensive coordination of gesture with protophones across the first year.

One might suggest that motoric development is simply faster in the vocal domain than in the domain of hand and arm movement. But this suggestion does not undercut the conclusions of the present work, because the earlier development of vocal capacities, which are known to be among humankind's most complex motoric capacities (see SM, 2.2), would require explanation of its own. In fact, motoric development of vocal capacities may have been naturally selected to occur early, precisely because of the importance of hominin infant vocalization as a signal of wellness across human evolution, a signal that could have been noticed even when caregivers were not looking, while gesture would have had no such advantage.

The relative tendency to communicate in the vocal domain compared to the gestural domain in early life is not only important in informing our understanding of the emergence of the speech capacity in modern human development, but it also offers insights into the likelihood that vocal communication predominated in the evolution of language. We reason, consistent with the evo-devo perspective of modern theoretical biology, that if language indeed originated from gestural use, gestural activity should have occurred to a far greater extent than we saw.

3. Frequencies and Functions of Vocalizations and Gestures in the Second Year of Life (Burkhardt-Reed & Oller, in preparation)

Abstract

The origin of language is being pursued in comparative studies of vocal and gestural communicative development in human infants/children. Speculations on the evolution of language have invoked comparisons across human and non-human primate communication. While there is widespread support for the claim that gesture plays a central, perhaps a predominant role in early language development and that gesture played the foundational role in language evolution, much empirical information does not accord with the gestural claims. The present study follows up on our prior work (Burkhardt-Reed et al., 2021) challenging the gestural theory of language development with longitudinal data showing that early speech-like vocalizations occurred more than 5 times as often as gestures in the first year of life. Now we bring longitudinal data on the second year (13, 16 and 20 mo.), showing again that vocalizations predominated, and especially in conventional (learned) communication, where 11 times more spoken words were observed than gestures that could be viewed as signs, thus equivalent to words. Our framework of observation highlights the fact that more than ¾ of gestures across these second-year data were deictics (e.g., pointing and reaching), acts that while significant in supporting the establishment of referential vocabulary in both spoken and signed languages, are not signs, but have single universal deictic functions in the here and now. In contrast, words and signs are functionally flexible, making possible reference to abstractions that are not bound to any particular illocutionary force nor to the here and now.

Introduction

Vocalization and Gesture in Communicative Evolution

Many have argued for a close relationship between gesture and language in terms of both evolution and development (Caselli et al., 2005; Iverson & Goldin-Meadow, 2005). A multitude of findings have been interpreted as showing that gestures are the first means to convey communicative intent prior to the onset of words around the first birthday (Bates et al., 1979; Capirci et al., 1996; Caselli et al., 2012). During the second year, published empirical research has tended to suggest children display a moderate preference for gestures over words (Behne et al., 2012; Cochet & Vauclair, 2010; Lüke et al., 2020; Rowe et al., 2008). In the context of these observations, it has come to be believed by many that infant gestures play a predominant role in the first human communication that presages language. But the matter remains debatable. Do gestures truly proceed vocalization in modern human development and the evolutionary origin of language? Or is vocalization more foundational for communication?

To the extent that the issue of language origins has been raised, most published opinions exploring evolutionary possibilities have leaned toward a gesture-first hypothesis (Arbib et al., 2008; De Stefani & De Marco, 2019; Hewes, 1973). Historically, studies that provide support for the gestural origins viewpoint suggest that the widely acknowledged vocal limitations of non-human primates, compared to their greater gestural flexibility, offer relevant evidence (Corballis, 2020; Hauser et al., 2014; Tomasello & Call, 2007); the advocates reason that since we are primates, our most fundamental communicative inclinations should be expected to resemble those of other primates. Furthermore, there has been speculation that adults are better equipped to interpret a baby's nonverbal communications than verbal expressions prior to the appearance of intelligible speech (Kendon, 2017). However, it is not obvious that this is true. The opinion

appears to depend on how one interprets the term “communication”. If, for example, communication is defined to be limited to acts that designate objects or other entities (deictics) and thus supply a basis for reference to objects or other entities in the here and now, then indeed, early infants are hampered in the vocal domain by not having words with which to make reference to objects or other entities. Pointing and reaching, on the other hand, are gestural acts that can serve the deictic function at least by late in the first year.

But if we define communication more broadly to include, for example, affective displays, it can be seen that vocalizations serve such functions from as soon as infants can breathe (Oller et al. 2019), not just in crying and whimpering, but also in vast numbers of speech-like vocalizations (“protophones”, Oller, 2000) that begin on the first day of life and greatly outnumber cries across the whole first year (Iyer et al. 2006). Furthermore both vocalizations and gestures can be viewed as fitness signals, especially if they are produced in comfort, and although infants do not have to intend such actions as communications, they may serve the function of communication of fitness even so, on occasions when caregivers notice them (Locke, 2006; Oller & Griebel, 2021).

One might imagine that the default hypothesis for those unfamiliar with the gesture-first literature would be that vocalization would have played the primary role in the origin of language, simply because language is primarily vocal. There are many categories of vocalization that human caregivers recognize in early communication development, including sounds not closely related to language such as crying and laughter, but several more termed, for example, “vowel-like” sounds (technically “vocants”), squeals, growls, and raspberries. Modern human infants display a drive to engage in vocalization, to listen to their own sounds and those of caregivers, and to pay attention to faces, especially the face of mothers from birth (Trevarthen,

1979; Trevarthen, 1988). It has been reasoned that human infants participate in vocal exploration, an activity that forms a foundation for language (Stark, 1980; Stark et al. 1993; Kent and Bauer, 1985).

Natural Selection of Communicative Signals

These tendencies of infants suggest humans have been naturally selected to display inclinations and capacities from infancy that promote communication with their caregivers and presumably later with all the members of their group. The patterns are compatible with the essentials of evolutionary developmental biology (evo-devo), a modern advance in biological theory (Carroll, 2005; Newman, 2016). Evo-devo proposes that major evolutionary changes have often involved selection on changes in rates and/or patterns of early development. Within the evo-devo perspective, it has been proposed that selection pressures may have been placed on the human infant to be inclined to seek and construct vocal capacities through vocal exploration (Oller, 2000; Oller and Griebel, 2021). Ultimately the exploration of vocalization in infancy forms a foundation for language, and the evo-devo perspective supports a constructionist view in which infants and young children are portrayed as building vocal language on their own (Tomasello, 2003).

The longitudinal observational literature on typical language acquisition indicates clear patterns in infants' early vocalizations and babbling, with an increase in not only the complexity of vocalization with age, but also of the degree to which pre-speech vocalizations resemble real words in natural languages (Oller, 1980; Roug et al., 1989; Stark, 1980). That is, infants build the capability to produce vocalization in a speech-like manner and also to use it communicatively, increasingly interacting in a conversational and expressive manner with caregivers and gradually incorporating real words into the interactions. Thus, the rate and complexity of vocalization

increases as children begin to use more fully semantic lexical items (words) even though they continue to produce non-speech protophones well into the second year (Eilers et al., 1993).

Evo-devo offers a framework for understanding how vocal language may have evolved. The development of protophones progresses toward more and more speech-like vocalization, and the high activity level of human vocalization compared to the much lower level in other apes (Hauser, 1996; Oller et al. 2019), suggests there must have been strong selection pressure on that high activity level throughout hominin history. The argument in favor of this idea invokes the notion that the hominin infant was born altricial (helpless) due to the narrowed hominin pelvis that accompanied bipedalism and thus required more long-term caregiver investment than in the cases of other ape infants (Locke, 2006; Oller and Griebel, 2021). Consequently, it is argued that the hominin infant was under selection pressure to produce fitness signals in the form of vocalizations suggestive of wellness, to augment the existing wellness indicators supplied by other factors such as normal motoric development, normal eating patterns, skin condition, and so on (Oller and Griebel, 2005).

Vocalization is particularly well suited to serving as a fitness signal, since caregivers do not have to be looking at infants in order to notice the extent to which protophones indicate comfort and/or well-being. Gestures on the other hand require caregivers to be looking, if the actions are to serve as fitness signals. 4000 species of songbirds produce massive amounts of fitness signaling in vocal mating displays that are broadcast even in the absence of any nearby potential mates (Kroodsma, 1999), but birds use such visual displays only when in the company of potential mating partners (Mitoyen et al. 2019). It is also notable that birds are born altricial, and that songbirds, like humans, produce a sort of babbling (subsong) in the fledgling stage

(Nootebohm, 1999), presumably both as a fitness signal and as a means of developing foundations for song, again in a manner that seems parallel to the human case.

Consequently it is plausible that the human line has long been under strong selection pressure to use vocalization as a fitness signal, while other primates, being less altricial at birth, have experienced less such pressure and thus have not developed such a strong vocal tendency. Gesture is available as a flexible means of communication for both humans and other apes, but the requirement of visibility of the gesturer offers reason to doubt that selection pressure would have influenced gesture to constitute such an important fitness signal as vocalization either in the origin of language or in the development of language; this reasoning is consistent with the observation that gestural activity occurs much less frequently than vocal activity throughout human life, except in cases of individuals who use sign language.

Nonetheless, there has been a great deal of attention paid to gesture in speculations about the origin of language. Theories of language origins in support of a vocal source are not absent, however (Seyfarth, 1987; Lemasson, 2011; Zuberbühler, 2017). In contrast to previously held beliefs about primarily gestural flexibility in non-human primates, some recent studies in support of vocal origins actually indicate more flexibility in the vocal than gestural communication systems of non-human primates (Byrne et al., 2017; Ey et al., 2009; Lameira et al., 2022).

Testing the Origins Question With Quantitative Developmental Evidence

Our evo-devo reasoning has always included the following supposition: If the earliest language was gestural, it would be reasonable to expect to see extensive gesture usage in early development. In turn, if the earliest language was vocal, it would be reasonable to expect to see extensive vocalization in early development. We pursued this reasoning in Burkhardt-Reed et al (2021), hereafter, “BR2021”.

A key factor in our research involved developing a framework of description that would allow justifiable quantitative comparisons of rates of vocalization and gesture in the first year. We view this as an important step, since according to recent reviews, the relative frequencies of human infant vocalizations and gestures as determined with clear definitions and criteria for classification have not been a major focus of empirical investigation on gesture and vocalization as language foundations in prior research (Prieur et al., 2020; Rodrigues et al., 2021). A key issue in developing an appropriate framework of description is recognizing the special role played by gestural transmission of deixis (e.g., pointing and reaching), a particular communicative function that is not easy to transmit in vocalization. These prelinguistic deictic gestures appear to be universal, having been documented across a variety of cultural groups, with essentially spontaneous emergence around 9 months (Cameron-Faulkner et al., 2020; Bates, 1976; Colonnesi et al., 2010; Özçalışkan et al., 2016). Studies have shown that children increase their deictic gesture use, specifically pointing, across the second year as a foundation for and an accompaniment to their emerging lexicon (Crais et al., 2004; Masur, 1983; Moreno-Núñez et al., 2020; Ramos-Cabo et al., 2019). One reason deictic gestures seem so important is that they can designate a clear referent, and discerning their message may thus seem less cognitively demanding than conventional gestures that have to be learned and may require more inference on the part of the observer (e.g., wiggling the body back and forth to indicate a worm) (Özçalışkan et al., 2018).

In BR2021 we argued that it is critical to recognize the distinction between the Universal deictic gestures as opposed to both Non-social gestures (e.g., rhythmic hand banging, opening and closing a hand as if practicing the movement) and Conventional gestures (such as clapping or waving hello). The BR2021 paper advanced the claim that these three types of gesture should

be sharply distinguished, providing a basis for comparison with three parallel kinds of vocalizations: Universal affective displays (such as crying and laughter), Non-social vocalizations (protophones that are not directed to others), and Conventional vocalizations (both performative and symbolic words). We specifically noted that the Universal types are neatly distinct across the modalities. Deixis can be performed with ease in the gestural domain, whereas affective communication cannot. Instead affect can be communicated by gesture only after lexicalization of gestures as signs. In contrast, affective communication occurs in vocalization from the first day of life (e.g., crying), but deixis can only be performed vocally after notable learning of speech ("look at the bunny that just came into my office"). Of course affect can be communicated by facial expressions from the first day of life.

In the present study, we continue the work of BR2021 into the second year of life where reaching and pointing are expected to play significant roles and some signed language symbology may also emerge, even in infants where sign language is not present in the household. At the same time, vocal communication is expected to grow rapidly. Thus, we aim to determine again the quantitative extent to which gesture and vocalization differ, this time in the second year of life. In addition, this effort affords the opportunity to add to evidence regarding the likelihood that language originated primarily in either the gestural or the vocal domain in our ancient ancestors.

Study Goals and Hypotheses

Our primary goal is to extend the research initiated in BR2021 into the second year, comparing voice and gesture rates by assessing non-social usage, lexicality, universality and directivity of vocal and gestural communicative behaviors. To explore meaningful patterns of similarity and distinction between gesture and vocalization, the present study observed typically

developing children at 13, 16, and 20 months in naturalistic laboratory recordings in accord with the following hypotheses:

(1) Prior research indicates higher rates of vocalizations compared to gestures in early communication development. Thus, we hypothesize that vocalizations will occur at a higher rate than gestures overall in the second year. More specifically:

   a. *Non-social* Vocalizations will occur more frequently than *Non-social* Gestures.

   b. *Universal* Vocalizations will occur more frequently than *Universal* Gestures.

   c. *Conventional* Vocalizations will occur more frequently than *Conventional* Gestures.

(2) In accord with current literature, gaze toward another person appears to occur more during vocalization in the first year than during gesture. But the pattern may change in the second year because as gestures (particularly pointing and reaching) become more prominent as intentional communications in the second year, the infant may need to ensure that caregivers are looking, in order to make the gestures communicatively effective. We project the following age effects in the second year:

   a. At 13 months vocalizations will occur more frequently with directed gaze than gestures.

   b. At 16 months vocalizations will occur more frequently with directed gaze than gestures.

   c. At 20 months gestures will occur more frequently with directed gaze than vocalizations.

Methods

Participants

Approval for the research was obtained from the IRB of the University of Memphis. Data were acquired from the University of Memphis Origin of Language Laboratories (OLL) archived longitudinal audio-video recordings. For the present study, we selected available recordings of 12 parent-infant pairs (6 male, 6 female) in the second year. The pairs were recorded while engaged in naturalistic play and interactions in the OLL. Families were recruited from child-birth education classes and by word of mouth to parents or prospective parents of newborn infants. Interested families completed a detailed informed consent indicating their interest and willingness to participate in a longitudinal study on infant sounds and parent-child interaction. All families lived in and around Memphis, Tennessee, and all infants were exposed to an English-only environment. Criteria for inclusion of infant participants included a lack of impairment of hearing, vision, language, or other developmental disorders. Demographics and recording ages for each infant at each recording session are provided in Table 5.

Table 5. Demographics and Recording Ages
This table displays recording ages and demographics in months and weeks at each session for each participant.

Infant	Gender	Age at Recordings (months; weeks)		
		13 mo.	16 mo.	20 mo.
1	F	13;1	16;2	20;1
2	M	13;1	16;0	20;1
3	M	13;1	16;0	20;0
4	F	13;3	16;3	20;3
5	F	13;2	16;2	21;1
6	F	13;1	16;0	20;3
7	F	13;0	16;0	20;0
8	M	13;0	16;1	20;1
9	F	13;1	16;2	20;2
10	M	13;3	16;1	20;3
11	M	13;0	16;0	20;0
12	M	13;1	16;1	20;1
Average age in months; weeks	*M* (*SD*)	13;1 (0;1)	16;1 (0;1)	20;2 (0;3)

Laboratory Recordings

Each of the longitudinal recordings included three sessions, each approximately 20-minutes in length, usually drawn from a continuous ~60-minute recording. For the current work, we analyzed sessions termed "interactive" at approximately 13, 16, and 20 months. In the other two sessions of the 60 min, the caregiver was in the room with the infant either reading in one case or talking to an interviewer in the other. From these recordings, one interactive session was selected for each infant (total: 12) at each age (total: 3), for a total of 36 recordings. Considerable amounts of vocalization (>4 protophones per min) occur in all three session types in OLL research (Oller, et al. 2021). We chose to analyze the interactive sessions hoping to maximize the amount of observable gesture.

The laboratory setting was designed to resemble a child's playroom equipped with eight cameras positioned in the corners of the room. Both parents and children were equipped with high fidelity microphones worn in an infant vest and on the parent's collar/shirt.

In the interactive sessions, the parents were instructed to interact naturally and playfully with the infant for the designated period of approximately 20 minutes. An experimenter in the adjacent control room selected two channels of video at each point in time. The cameras were switched as needed to obtain a view of the child's face as well as another view of the interaction between the child and the parent and/or researchers during the recording.

Coding

All the coding was conducted in the same software environment used in BR2021; AACT (Action Analysis Coding and Training, Delgado et al., 2010) is that software environment, allowing for simultaneous viewing of video/audio synchronized to frame accuracy. The real-time acoustic displays are provided in TF32 (Milenkovic, 2001). More detailed information about the coding software environment can be obtained from previous work (Oller et al. 2021).

The present research aimed to examine the frequency and directivity of communicative acts in the second year for both gesture and vocalization. Consequently, coding for the primary data collection was conducted in a way similar to that of BR2021, with both gestures and vocalizations coded on separate passes using repeat-observation coding (repeat viewing for gestural type and repeat listening for vocal type). This methodology allowed us to flexibly implement our descriptive framework for maximally meaningful comparison between vocal and gestural events in the second year. As in the prior study, we drew a distinction between actions and functions, where functions indicated potential communicative *intent* (designation, conversation, refusal, naming, etc.) The dimensions (or "fields") of coding were vocal acts,

gestural acts, gestural illocutionary functions, vocal illocutionary functions, gaze directivity for gesture, and gaze directivity for vocalization.

The first author was the primary coder. Before coding a segment, the primary coder viewed each ~20-minute recording to gain perspective on the flow of events from both infants and caregivers. Vocalizations and gestures were coded for each segment in two separate passes. In the first gesture pass, the coder used repeat-observation to designate occurrences and create boundaries for individual gestural acts. Repeat observation allows coders to determine onset and offset boundaries of each vocal and gestural event. AACT allows coders to create boundaries by placing a cursor at the start and a cursor at the end of an event. Then, a label for the event (or act, in this case) can be selected from a coding panel or by a designated keystroke. All vocal acts for each of the 12 participants had already been coded in a previous study. Consequently, we used the occurrences and boundaries from the previous vocal act coding for the present study.

Once an event is bounded, a placeholder can be created in a new dimension ("field") and the sequence designated by the placeholder can be selected and played repeatedly in AACT. In the second pass, the coder used the boundaries for each action in the vocal and gestural domain to automatically create placeholders for a new illocutionary dimension (for definitions of illocutionary forces, see Austin, 1962, BR2021 or Oller and Griebel, 2021), the categories of which indicated the social or non-social functions of the gesture or vocalization. The bounded time frames were transferred to a new coding panel without showing any vocal or gestural action codes, which allowed the primary coder to designate a new illocutionary function for each action in each modality. In the final step prior to analysis, we collapsed observed vocal and gestural functions into the three global function types, *Non-social, Universal, and Conventional,* to allow theoretically-motivated comparison across the modalities. BR2021. A full list of the illocutionary

functions utilized in the present work is provided along with the correspondence to gestural and vocal acts in Table 6 and Table 7 respectively.

Table 6. Gestural Illocutionary Functions

This list is intended to include illocutionary functions of gestures in infancy and early childhood that could conceivably be interpreted by any communicative partner or observer. The list includes the complete set used in this study. The terms are drawn partly from literature on ape and human infant gesture.

Function Category	Gestural Function	Definition
Non-Social	Non-social act	Any non-utilitarian act, not conveying communicative intent (e.g., rhythmic hand banging, body rocking)
Universal Social	Request object	Show desire to obtain something
	Accept	Receive something offered
	Offer	Present something to someone to accept/reject
	Request up	Show desire to be held or picked up (e.g., arms reaching up to indicate request)
	Designate	Indicate person or object of interest (e.g., pointing)
	Exult	Show happiness or excitement, celebrate (e.g., clapping hands together)
	Request help	Showing or bringing attention to a desired object or outcome (e.g., winding a toy)
	Refuse	Indicate unwillingness to do something (e.g., a hand block or turning away)
	Show	Make something visible to be perceived by another
	Social Playful	Engage in interactive exchange, usually involving an object (e.g., playing catch)
	Seek attention	Show desire to engage with another (e.g., touching or tapping a person to engage)
Conventional	Conventional gesture	Displaying performatives (e.g. blowing a kiss)
	Bye-bye	Wave “bye-bye”
	Greet	Wave “hello”
	Surprise	Showing excitement for an event or outcome (e.g., jack in the box)
	All gone	Indicating that something is completely finished or used up
	Imitation	Gestural imitation immediately or nearly so after another’s gestural production
	Signed name	Indicating reference to a particular object or person (e.g. signing for “hat” by patting head)
	Signed Comment	Indicating reference by providing information about the object or person
	Signed Sentence	More than one sign to indicate or convey information
	Continue	Responding to a conversation with gesture

Table 7. Vocal Illocutionary Functions
This table displays vocal illocutionary functions used in this study. As in Table 6, this table displays a complete list of illocutionary functions presumably interpretable by any communicative partner or observer.

Function Category	Gestural Function	Definition
Non-Social	No Force	No discernable illocutionary intent
	Vocal play	Playing with sound (e.g. canonical babbling)
Universal Social	Object-directed	Talking to an object
	Complain	Displaying a range of distress sounds to indicate discontent
	Exult	Displaying laughter or other high arousal sounds to indicate joy or excitement
	Call initiation	Call for the attention of another person to start interaction or communication
	Social Playful	Displaying characteristics of objects or animals (e.g., car sounds "vroom" or mooing for a cow) during playful interaction
	Continue	Continue a conversation without active engagement or elicitation (e.g., mom talks to me and I respond)
	Imitation	Sound imitation immediately or nearly so of another's vocalization
Conventional	Performative	Expression that serves as a performance of
	Request	Asking for something
	Offer	Presenting something to someone to accept/reject
	Refuse	Indicate or show unwillingness to do something
	Accept	Receiving something offered
	Designate	Indicate a particular person or thing to share interest with another person
	Show	Indicate a desire to share attention on an object
	Name	Naming an object or person
	Comment	Comment about an object, entity, quality, or situation
	Solicit	Indicating a desire to engage with or obtain something from someone
	Question	Requesting for information (e.g., "where is it?)
	Answer	Responding to a question posed by another person
	Agree	Indicating a person or object of interest
	Deny	Refusal to admit that something is true
	Vocal Sentence	Expressing a complete thought (e.g., "soft teddy" or "Oh, I see it!")
	Performative Request	Expressing desire for assistance (e.g., "help")
	Bye-bye	Indicating that someone or something is leaving or no longer in sight
	Uh-oh	Indicating a mistake or that something bad happened
	All gone	Indicating that something is completely finished or used up
	Wow	Expressing excitement and/or surprise
	No	Indicating unwillingness to do something or expression of dissent
	Yes	Indicating willingness to do something or expression of approval or agreement
	More	Indicating desire for an additional amount
	Hello	Expressing someone's or something's arrival

In a third coding pass, infant gaze directivity was coded during each vocalization and during each gesture, for all cases where at least one of the two video views allowed an infant gaze directivity judgement. Any instances where a judgment could not be made were coded as "can't see". The gaze coding determines whether a vocalization or a gesture is produced while looking at another person or while not looking at another person.

Global Types of Illocutionary Functions

Non-social Acts. Non-social acts have no inherent social communicative function, although they can be brought into the service of communication (especially through learned associations with meanings) because they can be produced voluntarily. *Non-social gestures* such as rhythmic hand banging, the most common Non-social gesture observed in our work (Ejiri, 1998; Iverson et al, 2007), can be viewed as comparable to protophones (especially reduplicated canonical babbling). However, such gestural acts have the potential to be used as intentional communications. *Non-social vocalizations* include all the non-word protophones (including both non-canonical and canonical babbling) produced, for example, in periods of vocal play. For the purposes of the present work we treated all protophones, even if they occurred during social interaction, as Non-social on the grounds that protophones appear not to be developed on the basis of caregiver input, but instead on the basis of endogenous infant vocal exploration (Oller and Griebel, 2021). Just as gestural acts such as opening and closing a hand or rhythmic hand banging can be adapted as learned conventional communications, so all the protophones (but especially the canonical syllables) can also be adapted as learned communications, especially as words.

Universal Acts. Of particular interest for our comparisons are *Universal* forms of vocalizations and gestures. Universal forms of gesture and vocalization are usually communicative but require no associative learning, and their intended functions are interpretable to potentially any communicative partner. *Universal gestures* include acts that appear to have inherently social communicative intent. These types of gestures have a heavily deictic function. The most common examples are pointing, showing, or reaching for an object that cannot be obtained by a child independently. Universal gestures are capable of transmitting certain

critically important communicative functions in the gestural domain such as refusal, request, and designation (pointing), and it appears they require no learning.

In the exclusively vocal domain, these deictic functions can only be transmitted through symbols/words ("look at bunny behind you" or "I don't want that"). In our previous work, we did not include *Universal vocalizations* (such as cry and laughter) in our counts because we did not fully recognize the importance of comparing Universal gestures with Universal vocalizations at the time. After reflecting on the uniqueness of the Universal gestures and the fact that they have essentially fixed functions (usually associated with designation), we were reminded that crying, whimpering, and laughter (which express negative or positive affect) also have essentially fixed functions of affect expression. And notably we were struck by the fact that in both the gestural and vocal domains, the universal functions were essentially disjunct in early infancy: the functions of deixis could be served gesturally but not vocally, while the functions of affect expression could be served vocally but not gesturally. Thus it became clear that one of the most interesting possible comparisons of gestural and vocal communication in infancy involved universal acts whose functions were markedly modality specific.

We did not make a direct comparison in BR2021 of the amounts of Universal gesture and Universal vocalization. Further, such a comparison has been ignored in all other prior attempts to quantify the relative amounts of gestural and vocal communication in infancy, as far as we know. The data from BR2021 showed growth in the amount of Universal gesture across the first year. In preparation for the present work we estimated from first year vocal coding for the 12 infants of the present study, that Universal vocalizations, crying and laughter, occurred more frequently in interactive sessions than Universal gestures at the early age (3 mo), but that rates for the two modalities were comparable beyond 6 mo. The clear trend in BR2021 showed Universal gestures

growing in frequency across the first year. There is reason to expect that trend to continue through the second year as infants increasingly point and reach communicatively both during purely gestural and combined vocal/gestural communications.

An additional universal tendency in human infants is vocal responsivity in face-to-face interaction using protophones, a pattern that is reported to begin at least to some extent from the first month of life (Trevarthen, 1979; Gratier et al. 2015). As the infant matures through the first year, it appears that this interactive tendency with non-word vocalizations grows. This tendency toward "protoconversation' is strong even though most of the vocalizations that occur during periods where caregivers attempt to elicit vocalization are not responsive (not conversational), but rather appear to constitute disengaged, endogenous vocal activity (Long et al., 2020). The fact, however, that many such vocalizations do indeed occur during engagement with caregivers without qualifying as conventional (learned) vocalizations poses an additional issue for our proposed categorization of vocalizations and gestures. It seems best to treat these conversational protophones as expressions of a special kind of (social) affect since they occur predominantly in affectively positive exchanges. Consequently, in the data presented below, conversational protophones will be treated as Universal vocalizations. It is notable that in a sense, gestures such as pointing, serve a dual function of deixis and visually-based interaction with caregivers; so both some of the Universal gestures and some of the Universal vocalizations we code can be viewed as having an important feature in common, namely that they both involve and promote engagement of infants with their caregivers.

Thus, we include Universal vocalizations in our proposed descriptive scheme in order to afford a comprehensive three-way comparison among communication types (i.e., Non-social, Universal, and Conventional) across the two modalities. Our rationale for advocating an explicit

comparison between the gestural universal deictics and the vocal universal vocalizations expressing affect is that without a lexicon, affect is not transmissible in the gestural domain, and designation (deixis) is not transmissible in vocalization, which makes this comparison imperative for our understanding of the relative ways children utilize universal communication in vocal and gestural language development.

Conventional Acts. *Conventional gestures* are those that are learned, and at the beginning, produced with a single discernible communicative (performative) function, such as waving "hello" or "bye-bye", or hand clapping in celebration. In our study of the second year, Conventional gestures also include signed words (e.g., patting the head with a flat hand with the palm facing down to produce the sign for "hat"), which are lexical items that can reach full semantic status as they develop, expressing different illocutionary functions on different occasions. We shall also keep track of signed sentences and conventional games involving gesture (e.g., peekaboo and patty cake). *Conventional vocalizations* also include performative words with a single illocutionary force, such as bye-bye, and fully semantic lexical items such as "hat", "doggie", or "mama/mommy" as well as sentences and vocal games.

Coder Agreement and Outcomes

Both the primary coder (first author) and the agreement coder (last author) had extensive experience in coding of infant vocal types (cry, laughter and both precanonical and canonical protophones), vocal illocutionary coding, infant gesture coding, gestural illocutionary coding, and gaze directivity coding. They designed the gesture coding scheme together over a period of three years of coding and recoding of audio-video samples prior to publications of BR2021. They recruited two additional members of the OLL coding team to participate in that activity. The results of the agreement tests (with correlations ranging from .7 to .9) for gestural, vocal and

directivity coding at the end of that three-year effort are reported in BR2021. Most importantly the pattern of coded results across the individuals on the segments selected for agreement tests were so similar that as stated in the article: "… if the entire data set had been coded by either of the agreement coders instead of the primary coder, none of the conclusions associated with the results reported [in the remainder of the article] would have changed". For the current effort the only difference in the coding task was that second year of life data were involved rather than first year of life data. The coding scheme, however, and the three global function categories remained the same as in BR2021.

For coding of the current data, the agreement coder followed the same coding procedure as the first author, using the criteria outlined in our previous study. Twenty-three segments from among the 36 recordings for the agreement study were semi-randomly selected with one five-minute segment selected from within each of 23 different recordings and coded by the agreement coder in a random order, thus mixing vocal and illocutionary fields as well as ages and infants. Seven or eight samples came from each of the three ages and all 12 infants were represented in the agreement samples. Five hundred thirty-five gestural and vocal events were coded by both coders of the agreement set. The agreement coder was blinded to the coding of the primary author.

The correlational results for the agreement coding showed high values. For Gestural illocution coding, there were 28 possible codes, 17 of which were actually used in the agreement coding. For the two coders across the 11 segments that were coded by both individuals the average correlation for number of gestures coded in each of the categories was $r = .95$, n = 11. For Vocal illocutionary coding, there were 36 possible codes, with again 17 vocal illocution codes that were actually used. The average correlation between the two coders was $r = .86$,

n = 12. Infant gaze directivity toward a caregiver occurred infrequently during both gestures and vocalizations. The agreement data were collapsed across the gestures and vocalizations so that all 23 segments were represented in the average correlation across the segments between the number of events that were coded as being directed toward a caregiver by gaze, $r = .94$, n = 23.

The overall patterns of coding by the two coders for vocalization and gesture with regard to the three global functional categories of events (Non-social, Universal, and Conventional) resembled each other substantially and also resembled the patterns for the entire data set as presented in the first section of Results, below. Figure 3 presents the comparison for the two coders. The patterns show considerable similarity in that both coders showed Universal gestures as by far the most frequent gestures and Conventional vocalizations being more frequent than Universal vocalizations, which in turn were more frequent than Non-social vocalizations. The difference between the coders on Universal vocalizations was accounted for predominantly by a tendency of the 2nd coder to categorize more vocalizations as conversational continuations than the 1st coder, who tended to treat the same utterances as Non-social. This is a common coding discrepancy because it is often ambiguous to observers even with audio and video whether an infant protophone is intended to be directed to a caregiver who is attempting to elicit conversation (Continue conversation, Universal vocalization) or whether the infant is in fact disengaged (Non-social).

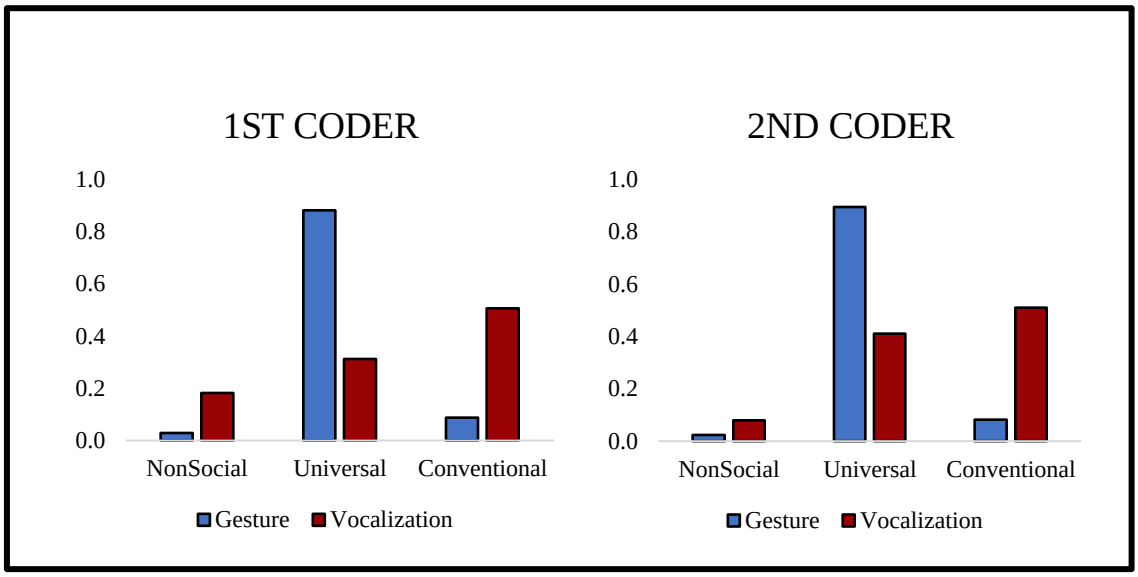

Figure 3. Coder Agreement Data
Proportion of events coded in the three global categories for the two individuals who independently coded 23 five-minute segments selected semi-randomly from 23 of the 36 recording sessions.

Data Analysis Plan

Our method is observational rather than experimental. Consequently we analyzed the data and present graphic representations where the primary issues at stake can be evaluated by overview of the figures. Still, the method allows formal statistical analysis. We have chosen to use Generalized Estimating Equations (GEE), a non-parametric approach, for most of the formal analyses since the study has a very small sample size with little likelihood to follow strong statistical assumptions and thus a non-parametric method is preferable (Liang & Zeger, 1986; Diggle et al., 2002). Also, GEE, unlike parametric approaches, requires no normality assumption, and the data in our study proved to show non-normal distributions. For the evaluation of the longitudinal hypotheses, therefore, we applied GEE analyses. Mann-Whitney U tests were utilized for formal comparison of gesture and vocalizations rates in the non-longitudinal summary data across the three global functional categories as presented in Figure 2.

Results

Overview

The research revealed more than twice as many vocalizations as gestures (3295 vs. 1455), but the differences were dissimilar across the three global types. For Non-social events, vocalizations were nearly 5.8 times more frequent (655 vs. 113), while for Universal events, gestures were 22% more frequent (1175 vs. 965). The greatest difference, as expected, was for Conventional events, where Conventional vocalizations were 10 times more frequent than Conventional gestures (1675 vs. 167). Figure 4 presents average amounts of each event type for the 12 infants with 95% confidence intervals as error bars. Vocalizations occurred most frequently as Conventional communications while gestures were produced most as Universal communications, with an average of fewer than 15 Non-social or Conventional gestures respectively for each infant, in contrast to nearly 100 Universal gestures per infant. Vocalizations, on the other hand, were plentiful in all three global categories, with 55 per infant for Non-Social, 80 per infant for Universal, and 140 per infant for Conventional vocalizations. Mann-Whitney U tests indicated, in spite of large inter-subject variance, that there were significantly more Non-social vocalizations ($p < .001$) than gestures and similarly that there were significantly more Conventional vocalizations ($p < .001$) than gestures.

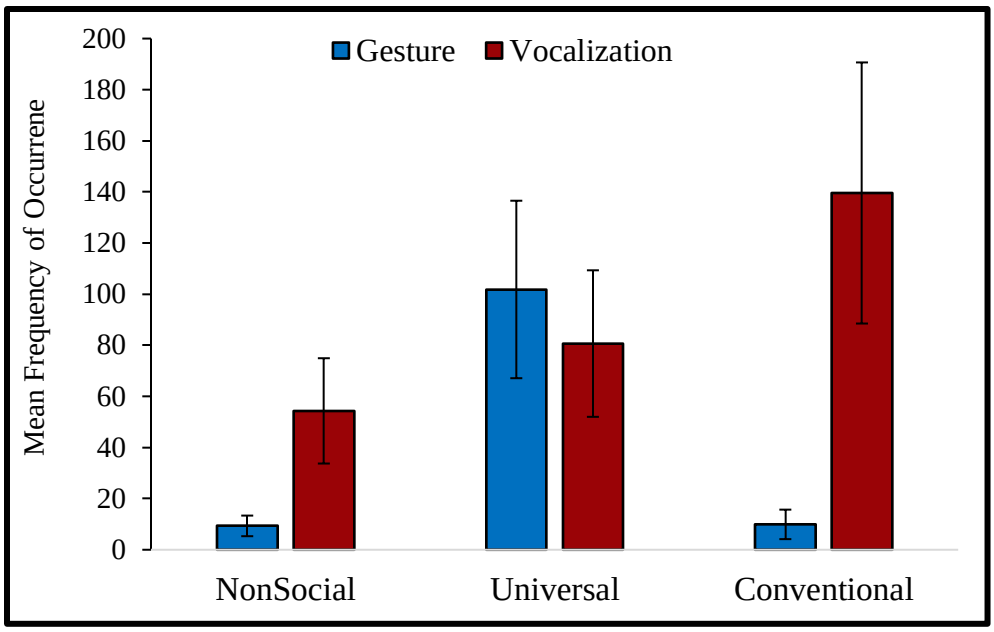

Figure 4. Frequency of Communicative Events per Infant Averaged across Ages
The data in the figure represent all 4750 gestural and vocal events and are based on means at the infant level. The error bars are 95% confidence intervals.

Hypothesis 1a.

The data relevant to Hypothesis 1a are presented in Figure 5, where it is clear that Non-social vocalizations were considerably more frequent at all three ages than Non-social gestures, confirming our expectations. The GEE analysis showed no significant interaction of Age with Modality, no significant main effect of Age, but a highly significant ($p < .001$) main effect of Modality. The GEE effect size estimate suggested there were 14.5 more vocalizations than gestures in the Non-social category per infant. Figure 3 is not based on the GEE modeled data but on the original infant-level data and the error bars are 95% confidence intervals. There were ~ 8 times more Non-social vocalizations than gestures at 13 and 20 months, and ~ 3 times as many at 16 mo.

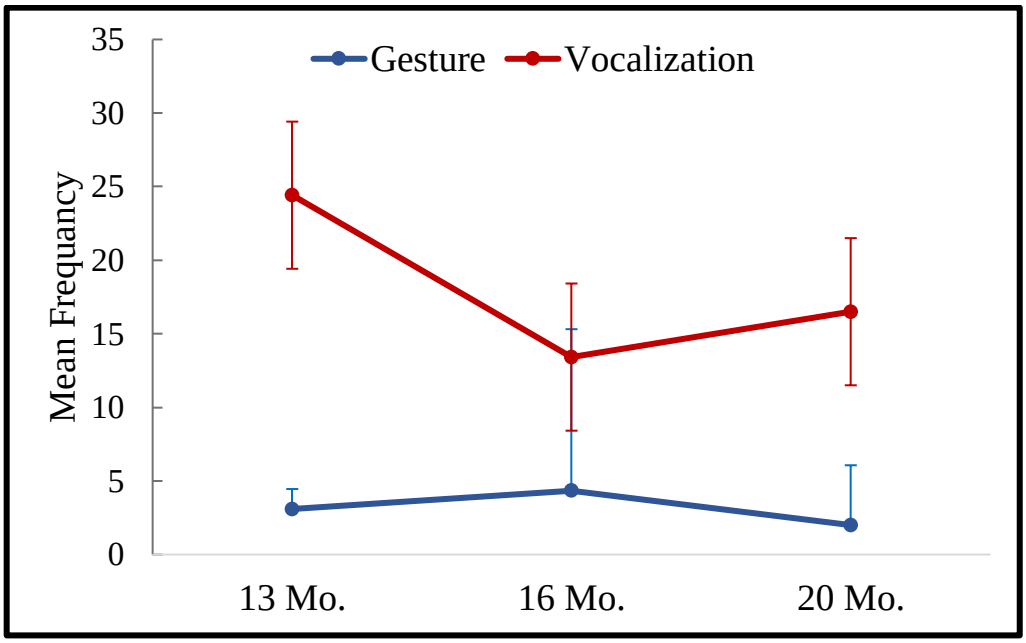

Figure 5. Mean Frequency per Infant of Nonsocial Acts
The data are based on the 768 gestural and vocal events that were coded as Non-social across three ages. The means and 95% confidence intervals are based on computation from the original data rather than from the modeled GEE values, which showed a statistically significant main effect of Modality, indicating higher rates of Non-social vocal than gestural events.

Hypothesis 1b.

Figure 6 shows a complex pattern of interaction of Age with Modality on Universal gestures and vocalizations. Hypothesis 1b was not confirmed, because there was a weak tendency for there to be more Universal gestures than vocalizations. The GEE analysis revealed a significant (p = .02) Age by Modality interaction, reflecting the sharp difference at 13 mo with respect to 20 mo, where the former showed a higher rate of Universal vocalization and the latter a higher rate of Universal gesture. Across the three ages, the Universal gesture rate rose while the Universal vocalization rate fell.

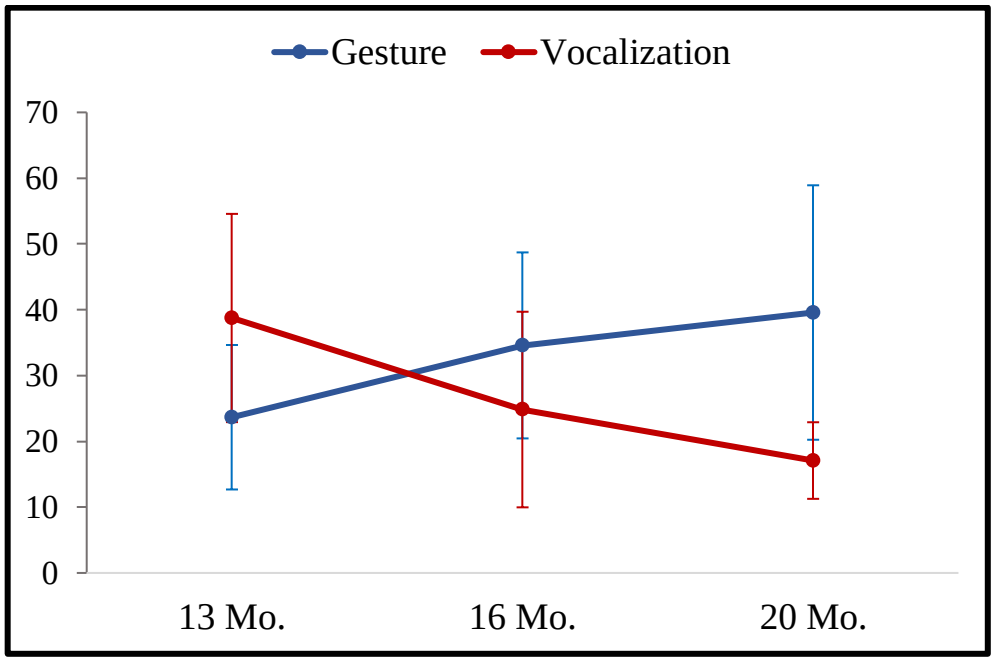

Figure 6. Mean Frequency per Infant of Universal Acts
The data are based on the 2140 gestural and vocal events that were coded as Universal communications across three ages. The means and 95% confidence intervals are based on computation from the original data rather than from the modeled GEE values, which indicated a significant interaction of Age by Modality, reflecting the rise in gestures across Age contrasting with the fall in vocalizations across Age.

Hypothesis 1c.

The strong tendency seen in Figure 7 for Conventional vocalizations (performative and symbolic words) to grow across Age is contrasted with a tendency for Conventional gestures (learned performative gestures and signs) to be very low across the entire Age range. This tendency was reflected in a highly significant Age by Modality interaction ($p < .001$), reflecting the fact that the difference between the rate of Conventional vocalizations and gestures grew dramatically from 13 to 20 mo. There were 4.6 times as many Conventional vocalizations as gestures at 13 mo, 8.3 times as many at 16 mo, and 21.4 times as many at 20 mo. Hypothesis 1c was thus strongly confirmed.

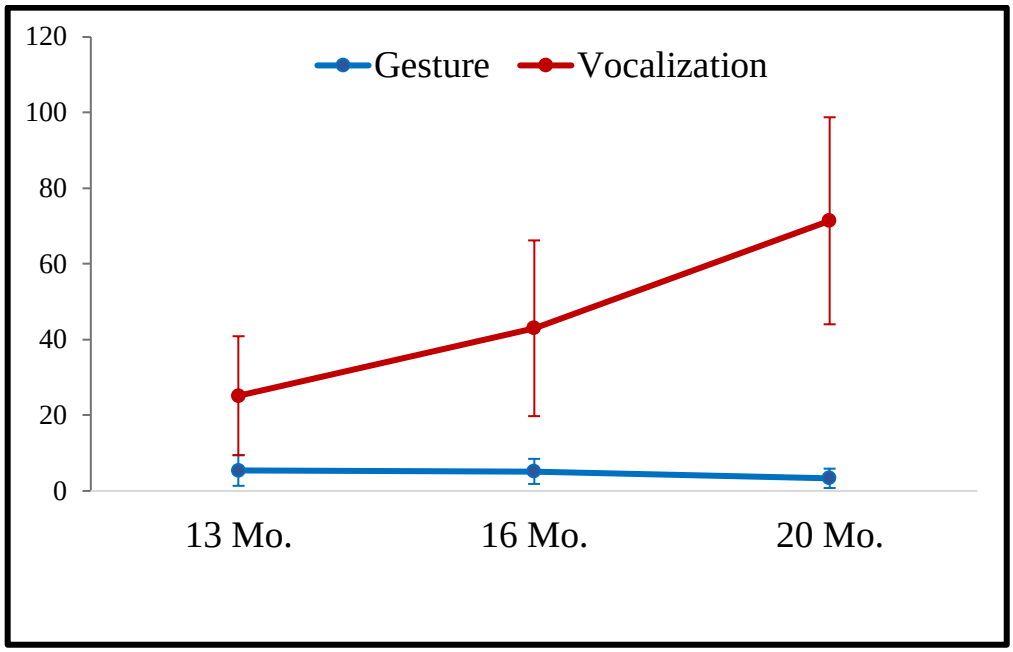

Figure 7. Mean Frequency per Infant of Conventional Acts
The data are based on the 1842 gestural and vocal events that were coded as Conventional communications across three ages. The means and 95% confidence intervals are based on computation from the original data rather than from the modeled GEE values, which indicated a significant interaction of Age by Modality, reflecting the rise in vocalizations across Age contrasting with no such rise in gestures across Age.

Hypothesis 2

The data on directivity are presented as pie charts in Figure 8, where the three rows show that our expectation for gesture to be more directed toward caregivers only at 20 mo was violated. In fact gestures were more directed toward caregivers at all three Ages. The GEE analysis revealed a main effect of Modality, $p = .003$, with no Age effect and no interaction of Age by Modality, even though the difference favoring gesture for directivity dropped to almost nil by 20 mo. The largest difference was at 13 mo, where gestures were accompanied by gaze directed toward a caregiver 33% of the time, while vocalizations were accompanied by directed gaze 26% of the time.

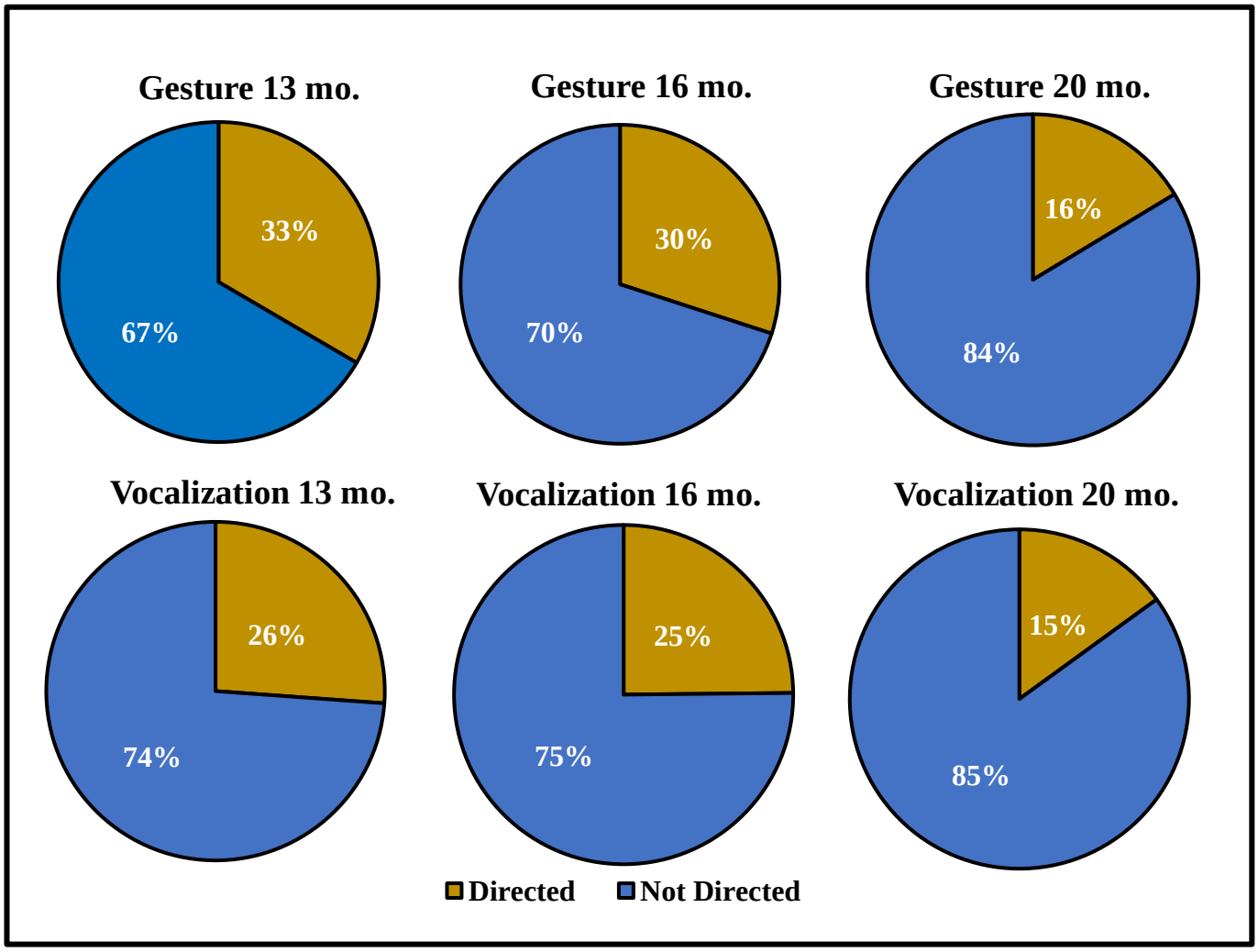

Figure 8. Proportions of Gaze Directed Gestural and Vocal Events
The data are based on all 4750 coded gestural and vocal events. Again the means reflect the raw data rather than the modeled GEE values.

Discussion

General Outcome Summary

Our intent is to shed light on the origin of language, especially with regard to the relative roles of vocalization and gesture both in evolution and in development. The research reported here has shown that, as predicted in our reasoning, the bulk of communicative activity in the second year, as in the first, was vocal rather than gestural in the infants we tracked. Overall more

than twice as many vocalizations as gestures were coded across the 36 twenty-minute samples of caregiver-infant interaction in the second year of life, a difference reminiscent of the fact that across the first year in BR2021, 5 times more vocalizations occurred than gestures and that even at 11 months more than twice as many vocalizations occurred as gestures. But as we have been at pains to explain, the most insightful way to portray the relative roles of gesture and vocalization in language development is not through an overall comparison of rates, but comparison of rates of communicative functions parceled into at least **three domains** reflecting different ways that gesture and vocalization can be used in infancy and also later in life.

Across the second year of life data, **Non-social** vocalizations, consisting of both precanonical protophones and canonical babbling, occurred 5.8 times more commonly than Non-social gestures, the kinds of actions that might be thought of as signed babbling (Meier, 2006; Meier & Willerman, 2013; Petitto & Marentette, 1991). In contrast, **Universal** gestures outnumbered Universal vocalizations by 21%, a fact that can be broken down into subcomponents: 93% of the Universal gestures were points or reaches (the latter of which were largely interpreted by the coder as requests) in about equal amounts, whereas the Universal vocalizations were composed of 44% positive or negative affect expressions (laughing, crying, whimpering) and 56% conversational continuations using non-words, i.e., protophones. The biggest difference for our three global categories was found in that **Conventional** vocalizations (both performative and symbolic words and/or sentences) occurred 10 times more frequently than Conventional gestures (performative or symbolic signs). This difference was greatly magnified at 20 months, where more than 21 times as many words occurred as signs (gestures that are functionally equivalent to words).

We conducted our study observing only circumstances where parents and infants were engaged in playful interaction. This approach was intended to maximize the amount of possible gestural interaction (although the amount of vocal interaction was also presumably enhanced), because a parent, even in the same room with an infant, but not engaged in communicating with the infant would likely not notice infant gestures. Vocalization produced in the same room has the advantage of being able to attract the attention of caregivers who are not communicatively engaged. It seems plausible that the greater amount of vocal than gestural communication in our study might be enhanced in the typical circumstance of the home, where caregivers have plenty to do other than engaging in playful interaction with their children.

Biological Perspectives on the Origin of Language

The outcome of our research is generally consistent with our perspective on the origin of language. We see considerable evidence that language is not only primarily vocal in mature users nowadays, but that communication is also primarily vocal at the beginning of human life, nowadays, with primary evidence being the reported quantitative differences. We further see reason based on these quantitative differences to support speculations that ancient hominins began in their evolution toward more complex communication than in other apes by evolving greater inclination and capacity for vocalization than in other apes (Hauser, 1996; Oller et al., 2019a). This evolutionary trajectory began, according to the reasoning, first with natural selection of exploratory infant vocalization as a fitness signal (Locke, 2006, 2009), which laid groundwork for later selection of more complex usages of vocalization to express a vast array of possibilities involving both indefinitely large numbers of illocutions and semantic units (typically words) as well as combinations of these making it possible to compose an indefinitely large set of possible sentences.

In principle, we can imagine that this kind of evolutionary trajectory might have begun with elaboration of gestural capabilities that were presumably already present in our distant ape ancestors. But the developmental evidence does not show the traces that we would expect if in fact gesture had formed the primary foundation for the massive differences that have been observed between the way humans communicate and the way our ape relatives communicate. Gesture occurs extremely infrequently in human infants in the first three or four months according to the observations of BR2021. Yet, humans show massive amounts of protophone vocalization (4-5 utterances per min every waking hour) during the same period, and the tendency to produce protophones at very high rates even begins two-months prior to due date in infants born prematurely and still in neonatal intensive care (Oller et al., 2019b). Our ape relatives, in contrast, show far less vocalization in infancy (Oller et al., 2019a), and the vocalizations they do produce have never been reported to be purely exploratory, while exploratory vocalization is the hallmark of human infant vocalization throughout the first year (Stark, 1981). So it makes sense to reason that a major transition must have occurred in ancient hominins, a transition where vocalization came to be produced far more frequently and far more flexibly than in other apes, and the fitness signaling theory suggests this transition occurred long before language (in any modern sense) existed.

Significant gestural usage does begin to occur more frequently in the second 6 mo of life than the first (BR2021), but still far less frequently than communicative vocalization. The present data confirm that the trend of increasing gestural usage (especially intentionally communicative gesture) also continues into the second year. Yet vocalization growth moves faster, and as we have seen, there is no point in time that we have observed, where gesture

actually provides the quantitatively primary form of human communication. Vocalization appears to dominate at every point.

Contrasting Approaches to the Study of Language Origins

So why does there exist a widespread belief that language originates developmentally in gesture? In the following section we explore ideas to help explain why our data appear to support primarily vocal origins of language and to contradict the widespread belief in primarily gestural origins. It appears that differences in how gestures and vocalizations are categorized in different realms of research probably form the primary basis for the differences in these perspectives.

Consider first the suggestion that gestures precede words in language development (Butcher & Goldin-Meadow, 2000; Iverson & Goldin-Meadow, 2005; Özçalışkan & Goldin-Meadow, 2005). Of course some gestural activity does precede the first words produced by infants, but it is also true that a great deal of vocal communication precedes the gestural activity. And our data suggest communicative vocalizations outnumber gestures at every age throughout early infancy and up through 20 months, where our current research ended. Thus we view the suggestion that gestures precede words as misleading because gesture do not precede vocal communication and never actually exceed it.

One needs, we think, to take a broader view of communication in order to assess the relative roles of gesture and vocalization in language development. We propose different criteria for categorization of infant actions than in many prior studies supporting gestural origins. In one clear case, a study sought to determine gestural roots of vocal communication and compared vocal word usage with "action gesture" usage (Caselli et al., 2012) based on a subtest from the MacArthur Bates Communicative Development Inventories (MB-CDI, Fenson et al., 2007). This methodology is based on parent report and provides no basis for differentiation among many of

the “Actions with Objects” on the list parents were instructed to check off in terms of ones that should be treated as gestures and ones that should be seen as purely Utilitarian. The method seems to us to have yielded a misleading impression, because, for example, in Caselli et al.’s Figure 1, it appears that gestures outnumbered words substantially throughout the 8-16 month period.

In our own categorization, “action gesture” events on the MB-CDI such as drinking from a cup, using a spoon, or opening a door are treated as *Utilitarian*, not gestural (see BR2021 for the technical definition). We agree with Caselli et al. that such actions are important developmentally, and they may even be important precursors to some aspects of communication. But they should not, in our opinion, be treated as gestures. There are additional steps that an infant would need to take in order to convert such actions into gestures. Suppose an infant picks up a cup and *pretends* to drink from it: We would treat such movements when they first appear as pretend play, because they implement a Utilitarian action (still not gestural, but closer to it than an act of real drinking) related directly to drinking from a cup. In a subsequent developmental step, such an action could become a sign with the meaning “drink” or “milk”, but first it would have to be shown that the infant had come to use the pretend-play-movement independent of any real or toy cup *as a gesture or sign* rather than as a Utilitarian act of imagination. In our view there must be a detachment of the physical action from the Utilitarian action.

The key point here is to draw the distinction between things we do with our hands because we need to eat, drink and open doors, for example, and things we do with our hands that are communications by design or purpose (true gestures). Our observations, because they are based on naturalistic audio-video recordings, put us in a position to draw distinctions among the

steps of development where a child's actions on an object or on an imagined object not actually in the child's hand can be made, and thus gestures can be differentiated from Utilitarian acts—the MB-CDI provides no such distinction. Consequently, we view the problem with prior literature using "action gestures" as Caselli et al. did as truly one of categorization.

A more direct comparison would ask how many words and how many gestural signs occur at various points in development. We know of no prior literature that has actually counted the number of times in a given period a child produces a word as opposed to a gestural sign, nor has any prior work assessed the communicative intent of observed events within such a comparison. Our report indicates there were massively more words than signs in the 36 caregiver-infant interactions.

Perhaps even more important to illustrating our objections to much of the literature under discussion, gestural-origin advocates have routinely ignored vocalizations that are not yet words, assuming that gesture is a precursor to language but that vocalization (prior to the development of words) is not. For instance, protophones (non-word vocalizations) are not counted at all by Caselli et al. Yet protophones are clearly significantly communicative, not just as fitness signals, but as indicators of infant state, and perhaps most importantly as mechanisms by which infants and parents engage in social interaction, the very embodiment of primary intersubjectivity (Trevarthen, 2001). It seems clear that face-to-face vocal interaction (often termed protoconversation) beginning in the first months of life is a necessary foundation for secondary intersubjectivity (joint attention), which in turn appears to be a necessary foundation for word and sign learning (Terrace, 2022; Tomasello, 2010; Oller et al., 2016).

We do not claim gesture is unimportant in language and language development. Especially we support the logically well-founded claim that joint attention plays a crucial

grounding role in the learning of substantive vocabulary (Bruner, 1975). And we have no doubt that pointing is an important method helping to establish joint attention in infancy and early childhood (Tomasello et al. 2007; Tomasello & Farrar, 1986). Other research also suggests that pointing both by infants and by caregivers can support word learning in the second year of life (Lucca & Wilbourn, 2018) and perhaps especially so as the end of the second year approaches (Paulus & Fikkert, 2014).

But our categorization of pointing (and reaching, to the extent that reaching is viewed as an act of joint attention) as a deictic event, with only one Universal communicative function (to designate an object or entity in the here and now), undercuts (or at least severely limits) the suggestion that a point plus a word is akin to a sentence (Iverson & Goldin-Meadow, 2005). A point is not itself a symbol, and cannot become one, unless we are willing to give up its deictic function, something that might be done for the sake of argument, to illustrate the flexibility of human learning, but it would not be a good idea to divest pointing of its natural function. The importance of pointing as a mechanism of establishing and developing joint frames of reference (Tomasello et al., 2007) does not promote it out of the status of an illocutionarily fixed action (the function being designation) into the realm of symbolism, where words and signs must be illocutionarily free to transmit a vast array of functions (naming, denying, correcting, requesting, insulting...) or else they would not qualify as words or signs. We do not intend to diminish the importance of pointing and naming or pointing and producing words, but it is confounding to interpret pointing as if it were a word or even as a sentential frame.

In BR2021 we noted that at 11 months the infants we studied showed remarkably little pointing, < 2% of observed gestures, whereas in the present data from the second year of life, pointing accounted for 41% of gestures over the whole second year and 59% at 20 mo. Much of

this tendency appeared to be associated with picture-book naming. The caregivers often chose a book from among the possible items for interaction (several options were always available) and engaged children in a naming game, and this tendency was especially strong at the older ages. The use of pointing appears to be very circumstance specific, and it is hard to know how often caregivers engage infants in the home in naming games where pointing plays a major role. Of course, our evidence is consistent with the idea that early word learning can be supported by pointing—indeed word usage in our data increased by a factor of 2.8 from 13 to 20 mo., during which period pointing increased by a factor of 5.4. But which determined which? Perhaps pointing drives word learning, but it might also be claimed that pointing is supported by word usage, and indeed much of the pointing we observed occurred in circumstances where caregivers named a picture in a child's book ("where is the goat", for example), and the child followed by naming the picture and/or pointing to it.

Gestural interaction with caregivers in our data was overwhelmingly associated with pointing, reaching, and other deixis-oriented Universal gestures. We observed no cases where gestures that might be viewed as sign babbling (Non-social gestures) were adapted to face-to-face interaction, while protophones were routinely utilized in face-to-face protoconversation. So all in all, it appears that Caselli et al., and presumably others (such as Iverson and Goldin-Meadow) supporting gestural origins of language based on developmental evidence have counted many items as gestures that we think should not be so counted, and have neglected to count as vocal communications the great bulk of the vocalizations that occur across the first year and a substantial proportion that occur in the second.

Summary on Gaze Directivity

Our data on gaze directivity provide food for thought. It was clear from BR2021 that gaze directivity was limited in the first year both for vocal and gestural actions, although the proportion of vocal events that were directed to a caregiver was higher than for gesture. We interpreted this pattern as indicating that the gestural material we witnessed in the first year was only rarely intended as communication, since a gesture when no one is looking can hardly be communicative. Vocalization in contrast just needs to be loud enough to be heard to be communicative, one of the key facts that has always been adduced to support vocal origins of language. Yet, once a communicative frame is established or once it is merely understood by two parties that an interaction between them is underway, gaze directivity is not actually necessary for either vocalization or gesture. The recordings we studied were overwhelmingly defined by precisely this sort of clear interactive frame. Both parent and infant knew they were interacting. Consequently perhaps, they even more rarely used directed gaze in the second year of life data than similar infants had in the first year of life (BR2021). In fact, gestures were slightly more commonly gaze directed than vocalizations in the second year data, with the amount of gaze directivity falling to less than 20% of both gestural and vocal events by 20 months. The communications in both modalities could in the great majority of circumstances be interpreted as intentional communications regardless of the lack of eye contact.

Limitations

In this section, we discuss provisos about the generalizability of the evidence we have presented here and in BR2021. First, our evidence in these two studies pertains to development only, and our reasoning about evolution of language requires reference to very different bodies of research. Second, our developmental evidence has limits of generalizability because in

households where deafness is involved among immediate family members, and thus where sign language may play a major role in communication, the patterns of usage of vocalization and gesture may be different from those we found in BR2021 and in the present work with families where all participants (infants, siblings, and caregivers) were hearing people. In households where sign language plays a major role, much of our perspective, based on the present research might need to be substantially modified, and we have developed no empirical basis for that modification. Third, there may be notable cultural differences in how much gesture is utilized both in infancy and adulthood, even in hearing households. In particular, evidence has been presented to suggest that Italian is a language where much more gesture is utilized than in other European languages, and that infants learning the Italian language may also use gesture more frequently than infants learning other languages (Iverson et al., 2008).

Still, the research we have conducted is surely relevant to the question of relative roles of vocalization and gesture in the origin of language. This contention is based in part on an evolutionary developmental (evo-devo) perspective (Arthur, 2010, 2021), where evidence has been mounting for decades indicating that major innovations in evolution typically require selection on developmental processes. If a major transition in evolution is to occur, it will typically not only involve changes in development, but traces of the evolutionary process will be left in development. We have expanded this reasoning to encompass the case of language evolution, by suggesting that the relative roles of vocalization and gesture in modern human development, both in quantity of events and in order of appearance of events may well reflect relative importance of vocalization and gesture in the process of emergence.

Clinical Implications and Future Directions

Our data highlight a favoring of the vocal system in human infancy and a key supporting role of gesture in language development. Although the evidence from the current study and BR2021 provide new perspectives on the emergence of language, future longitudinal studies may benefit from larger sample sizes and long-term naturalistic observation of samples of younger infants across the first two years of life, an adjustment which may serve to improve the generalizability of the research findings.

As for clinical implications, the results suggest that both gesture and vocalization could be targeted for early clinical intervention. If we see signs of deviation from typical development, as for example in emergent autism or specific language impairment, it is pertinent to determine whether interventions should be based primarily in vocalization, gesture, or both. For example, pointing, clearly a kind of gesture, develops in the second half year of life. Aberrations in pointing, especially very delayed development or absence of pointing, are strong risk signs for autism (Adamson et al., 2009; Mundy et al., 1990). Consequently, the study of early gesture, including possible precursors to pointing, has been an appropriate target for early detection research in this population. Understanding disorder-specific patterns, such as those in autism, can help guide clinicians in planning assessment procedures and treatment approaches that consider emerging gestures as well as spoken language. Thus, future directions of this line of research might examine differences in children with developmental disorders compared to trajectories in typical development. Studies of this nature may reveal unique patterns of development and use in different clinical populations, which could help inform clinicians on how to support children with early language and communication difficulties.

At the same time, the results of our work suggest that one could easily overplay the role of gesture in language development on the basis of the widespread claim that "gesture paves the way to language." Our results show clearly that vocalization occurs far more frequently than gesture and that it plays a fundamental and very frequent role in the development of social interaction.

4. Conclusion

This dissertation examined the frequency of occurrence of vocal and gestural events across the first two years of life. To evaluate the relative roles most appropriately we advocated a categorization scheme that assigns gestures and vocalizations to three global function types: 1) Non-social acts, either gestural or vocal, are not intentionally communicative, but are produced in such a way that they could potentially be associated with some communicative function through learning; 2) Universal acts are not learned, but are communicative by nature; and 3) Conventional acts involve learned associations of particular acts with particular illocutionary functions (in the case of performatives, such as "bye-bye") or learned associations of particular acts with semantic units (in the case of true symbols, such as a word or a sign referring for example to the class of goats, and being then adaptable for use with an indefinite number of illocutionary functions where the word goat could be involved).

We have ruled out as irrelevant to comparisons of rate of gesture and vocalization, acts that can be termed Utilitarian. Some have argued that "action-gestures" such as drinking from a cup should be treated as gestures. We, on the other hand, view the term "action-gestures" as a misnomer, implying falsely that an action such as drinking from a cup is analogous to an action that could be a sign—instead such acts are viewed in our own scheme as Utilitarian and are not entered into the comparison with vocal acts of communication or vocal acts that could be brought to the service of vocal communication through learning. We also rule out as irrelevant to comparisons of rate of gesture and vocalization any Utilitarian vocalization. Such vocalizations serve functions related to digestion (burping, for example) and respiration (hiccoughs, coughing, sneezing, for example) and are not thus available as potential learned acts of communication.

In Study 1, we examined the frequency of protophones and gestures across the first year, categorizing actions in terms of the three-fold scheme of global communicative functions. The findings indicated that protophones occurred at a significantly higher rate than gestures during the first year with regard to all three of the global functions. The results contradict the widespread belief that early language is founded primarily in gesture, which has been propagated by claims that gestures precede words in the development of language and thus likely in the origin of language.

In Study 2, we evaluated the rates of vocalization and gesture in the second year. Vocal communications occurred far more frequently in the second year, as in the first, than gestural communications. The results of both works support the idea that the primary foundations of language are vocal, even though the results also suggest an important role for gesture in the emergence of language. Having broken the gestural and vocal events into these three groupings, we found that vocalizations were very dominant in frequency of occurrence as Non-social and Conventional acts, but that Universal gestures and vocalizations occurred at similar frequencies, with gestures taking the lead by 20 mo. The gestures that were most common at 20 months were of the deictic type, the very type that persists even into adulthood as a Universal way of designating topics of reference in human interaction.

The predominance of vocalization in the overall amount of communication in both infancy and maturity is reflective of the fact that human languages primarily use vocal means of transmitting specific references to the present, the past, the future and the imagination, with gestures offering important though secondary elaborations on the vocal messages (McNeill, 1992). Of course fully semantic references and sentences can also be formed in signed languages, but our observations suggest there exists a bias in humanity favoring a primarily

vocal system, a bias that is seen at least from the first month through most of the second year of life.

References

Adamson, L. B., Bakeman, R., Deckner, D. F., & Romski, M. (2009). Joint engagement and the emergence of language in children with autism and Down syndrome. *Journal of Autism and Developmental Disorders*, *39*, 84-96.

Arbib, M. A., Liebal, K., & Pika, S. (2008). Primate Vocalization, Gesture, and the Evolution of Human Language. *Current Anthropology*, 1053-1076.

Armstrong, D. F., & Wilcox, S. E. (2007). *The gestural origin of language*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195163483.001.0001

Arthur, W. (2010). *Evolution: A developmental approach*. John Wiley & Sons.

Arthur, W. (2021). *Understanding Evo-Devo*. Cambridge University Press.

Austin, J. L. (1962). How to do things with words. Harvard University William James lectures 1955. Oxford University Press.

Bates, E. (1976). *Language in context*. New York: Academic Press.

Bates, E., Benigni, L. Bretherton, 1., Camaioni, L., & Volterra, V. (1979). *The emergence of symbols: Cognition and communication in infancy*. Academic Press.

Behne, T., Liszkowski, U., Carpenter, M., & Tomasello, M. (2012). Twelve-month-olds' comprehension and production of pointing. *The British Journal of Developmental Psychology*, *30*(Pt 3), 359–375. https://doi.org/10.1111/j.2044-835X.2011.02043.x

Bonvillian, J. D., & Patterson, F. G. (1999). *Early sign-language acquisition: Comparisons between children and gorillas*. In S.T. Parker, R.W. Mitchell, & H. L. Miles (Eds), *The mentalities of gorillas and orangutans: Comparative perspectives* (p. 240-264). Cambridge University Press. https://doi.org/10.1017/CBO9780511542305

Bruner, J. (1975). The ontogenesis of speech acts. *Journal of Child Language, 2*, 1-19.

Buder, E. H., Chorna, L. B., Oller, D. K., & Robinson, R. B. (2008). Vibratory regime classification of infant phonation. *Journal of Voice, 22*(5), 553-564. http://doi.org/10.1016/j.jvoice.2006.12.009

Butcher, C., & Goldin-Meadow, S. (2000). 12 Gesture and the transition from one-to two-word speech: when hand and mouth come together. *Lang Gest*, *2*, 235.

Burkhardt-Reed, M. M., Long, H. L., Bowman, D. D., Bene, E. R., & Oller, D. K. (2021). The origin of language and relative roles of voice and gesture in early communication development. *Infant behavior & development*, *65*, 101648. https://doi.org/10.1016/j.infbeh.2021.101648

Byrne, R. W., Cartmill, E., Genty, E., Graham, K. E., Hobaiter, C., & Tanner, J. (2017). Great ape gestures: Intentional communication with a rich set of innate signals. *Animal Cognition*, *20*(4), 755–769. https://doi.org/10.1007/s10071-017-1096-4

Call, J., & Tomasello, M. (Eds.). (2007). *The gestural communication of apes and monkeys*. Taylor &Francis Group/Lawrence Erlbaum Associates.

Carroll, S. B. (2005). Evolution at two levels: On genes and form. *PLoS Biol*, *3*(7), e245. http://doi.org/10.1371/journal.pbio.0030245

Caselli, M. C., Rinaldi, P., Stefanini, S., & Volterra, V. (2012). Early action and gesture "vocabulary" and its relation with word comprehension and production. *Child Dev*, *83*(2), 526-542. doi:10.1111/j.1467-8624.2011.01727.x

Cheney, D. L., & Seyfarth, R. M. (2005). Constraints and preadaptations in the earliest stages of language evolution. *The Linguistic Review*, *22*(2-4), 135-159. https://doi.org/10.1515/tlir.2005.22.2-4.135

Clay, Z., & Zuberbühler, K. (2009). Food-associated calling sequences in bonobos. *Animal Behaviour*, *77*, 1387–1396. https://doi.org/10.1016/j.anbehav.2009.02.016

Cochet, H., & Vauclair, J. (2010). Pointing gestures produced by toddlers from 15 to 30 months: Different functions, hand shapes and laterality patterns. *Infant Behavior and Development*, *33*(4), 431–441.

Colonnesi, C., Stams, G. J. J., Koster, I., & Noom, M. J. (2010). The relation between pointing And language development: A meta-analysis. *Developmental Review*, *30*(4), 352–366.

Condillac, E. B. d. (1756). *An essay on the origin of human knowledge; being a supplement to Mr. Locke's Essay on the human understanding (Translation of Essai sur l'origine des connaissances humaines)*. J. Nourse.

Corballis, M. C. (2010). The gestural origins of language. *Wiley Interdisciplinary Reviews: Cognitive Science*, *1*(1), 2–7. http://doi.org/10.1002/wcs.2

Corballis, M. C. (2020). Crossing the Rubicon: Behaviorism, Language, and Evolutionary Continuity. *Frontiers in Psychology*, *11*, 653. https://doi.org/10.3389/fpsyg.2020.00653

Crais, E., Douglas, D. D., & Campbell, C. C. (2004). *The intersection of the development of gestures and intentionality*. Journal of Speech Language and Hearing Research, *47*(3), 678-694. doi: 10.1044/1092-4388(2004/052

Delgado, R.E., Buder, E.H., & Oller, D.K. (2010). AACT (Action Analysis Coding and Training). *Intelligent Hearing Systems, Miami, FL*.

Diggle, P. J., Heagerty, P., Liang, K.-Y., & Zeger, S. (2002). *Analysis of Longitudinal Data. Oxford Statistical Science Series. Oxford Statistical Science Series*.

Donnellan, E., Bannard, C., McGillion, M. L., Slocombe, K. E., & Matthews, D. (2020). Infants' intentionally communicative vocalizations elicit responses from caregivers and are the best predictors of the transition to language. *Developmental Science*, *23*(1), e12843. https://doi.org/10.1111/desc.12843

De Stefani, E., & De Marco, D. (2019). Language, Gesture, and Emotional Communication: An Embodied View of Social Interaction. *Frontiers in Psychology*, *10*. https://www.frontiersin.org/articles/10.3389/fpsyg.2019.02063

Ejiri, K. (1998). Relationship between rhythmic behavior and canonical babbling in infant vocal development. *Phonetica*, *55*(4), 226-237.

Eilers, R. E., Oller, D. K., Levine, S., Basinger, D., Lynch, M. P., & Urbano, R. (1993). The role of prematurity and socioeconomic status in the onset of canonical babbling in infants. *Infant Behavior and Development*, *16*, 297-315.

Ey, E., Rahn, C., Hammerschmidt, K., & Fischer, J. (2009). Wild female olive baboons adapt Their grunt vocalizations to environmental conditions. *Ethology*, *115*(5), 493–503.

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. ed.). Paul H. Brookes Publishing Co.

Fogel, A., & Hannan, T. E. (1985). Manual actions of nine-to fifteen-week-old human infants during face-to-face interaction with their mothers. *Child Development*, 1271-1279.

Franco, F., Perucchini, P., & March, B. (2009). Is infant initiation of joint attention by pointing affected by type of interaction? *Social Development, 18*(1), 51-76. https://doi.org/10.1111/j.1467-9507.2008.00464.x

Gardner, R. A., & Gardner, B. T. (1969). Teaching sign language to a chimpanzee. *Science, 165*(3894), 664-672. http://doi.org/10.1126/science.165.3894.664

Gillespie-Lynch, K., Greenfield, P. M., Feng, Y., Savage-Rumbaugh, S., & Lyn, H. (2013). A cross-species study of gesture and its role in symbolic development: Implications for the gestural theory of language evolution. *Frontiers in Psychology*, *4*, 160. https://doi.org/10.3389/fpsyg.2013.00160

Gratier, M., Devouche, E., Guellai, B., Infanti, R., EbruYilmaz, & Parlato-Oliveira, E. (2015). Early development of turn-taking in vocal interaction between mothers and infants. *Frontiers in Psychology*, *6*, 1-10. https://doi.org/doi:10.3389/fpsyg.2015.01167

Gros-Louis, J., & Wu, Z. (2012). Twelve-month-olds' vocal production during pointing in

naturalistic interactions: Sensitivity to parents' attention and responses. *Infant Behavior and Development*, *35*(4), 773. https://psycnet.apa.org/doi/10.1016/j.infbeh.2012.07.016

Hauser, M. (1996). *The evolution of communication*. MIT.

Hauser, M. D., Yang, C., Berwick, R. C., Tattersall, I., Ryan, M. J., Watumull, J., Chomsky, N., & Lewontin, R. C. (2014). The mystery of language evolution. *Frontiers in Psychology*, *5*. https://www.frontiersin.org/articles/10.3389/fpsyg.2014.00401

Hewes, G. W. (1973). Primate communication and the gestural origin of language. *Current Anthropology*, *14*(1-2), 5-24. https://doi.org/10.1086/201401

Iverson, J. M., Tencer, H. L., Lany, J., & Goldin-Meadow, S. (2000). The relation between gesture and speech in congenitally blind and sighted language-learners. *Journal of Nonverbal Behavior*, *24*(2), 105-130.

Iverson, J. M., & Goldin-Meadow, S. (2005). Gesture paves the way for language development. *Psychological Science*, *16*(5), 367–371. http://doi.org/10.1111/j.0956-7976.2005.01542.x

Iverson, J. M., Capirci, O., Volterra, V., & Goldin-Meadow, S. (2008). Learning to talk in a gesture-rich world: Early communication in Italian vs. American children. *First Language*, *28*(2), 164-181.

Iverson, J. M., & Wozniak, R. H. (2016). Transitions to intentional and symbolic communication in typical development and in autism spectrum disorder. In D. Keen, H. Meadan, N. C. Brady, & J. W. Halle (Eds.), *Prelinguistic and minimally verbal communicators on the autism spectrum* (p. 51–72). Springer Science + Business Media. https://doi.org/10.1007/978-981-10-0713-2_4

Iyer, S. N., Ertmer, D. J., & Stark, R. E. (2006). Assessing vocal development in infants and toddlers. *Clinical Linguistics & Phonetics*, *20*(5), 351–369.

Iyer, S. N., & Oller, D. K. (2008). Prelinguistic vocal development in infants with typical hearing and infants with severe-to-profound hearing loss. *The Volta Review*, *108*(2), 115.

Iyer, S. N., Denson, H., Lazar, N., & Oller, D. K. (2016). Volubility of the human infant: Effects of parental interaction (or lack of it). *Clinical Linguistics & Phonetics*, *30*(6), 470–488. http://doi.org/10.3109/02699206.2016.1147082

Jhang, Y., & Oller, D. K. (2017). Emergence of functional flexibility in infant vocalizations of the first 3 months. *Frontiers in Psychology*, *8*, 300. https://doi.org/10.3389/fpsyg.2017.00300

Jhang, Y., Franklin, B., Ramsdell-Hudock, H. L., Oller, D. K. (2017). Differing roles of the face and voice in early human communication: Roots of language in multimodal expression. *Frontiers in Communication*, *2*(10), 1-12. https://doi.org/10.3389/fcomm.2017.00010

Kendon, A. (2017). Reflections on the “gesture-first” hypothesis of language origins. *Psychonomic Bulletin & Review, 24*(1), 163-170. https://doi.org/10.3758/s13423-016-1117-3

Kent, R., & Bauer, H. R. (1985). Vocalizations of one-year-olds. *Journal of Child Language, 12*, 491-526.

Kersken, V., Gómez, J.-C., Liszkowski, U., Soldati, A., & Hobaiter, C. (2019). A gestural Repertoire of 1- to 2-year-old human children: in search of the ape gestures [journal article]. *Animal Cognition, 22*(4), 577-595. https://doi.org/10.1007/s10071-018-1213-z

Koopsman-van Beinum, F. J., & Van der Stelt, J. M. (1986). Early stages in the development of speech movements. In B. Lindblom, & R. Zetterstrom (Eds.), *Precursors of early speech* (p. 37-50). Wenner-Gren Center International Symposium Series.

Kroodsma, D. E. (1999). Making ecological sense of song development. In M. D. Hauser & M. Konishi (Eds.), *The design of animal communication* (pp. 319-342). MIT Press.

Lameira, A. R. (2017). Bidding evidence for primate vocal learning and the cultural substrates for speech evolution. *Neuroscience and Biobehavioral Reviews, 83*, 429–439. https://doi.org/10.1016/j.neubiorev.2017.09.021

Lameira, A. R., Santamaría-Bonfil, G., Galeone, D., Gamba, M., Hardus, M. E., Knott, C. D., Morrogh-Bernard, H., Nowak, M. G., Campbell-Smith, G., & Wich, S. A. (2022). Sociality predicts orangutan vocal phenotype. *Nature Ecology & Evolution*, *6*(5), Article 5. https://doi.org/10.1038/s41559-022-01689-z

Lemasson, A. (2011). What can forest guenons "tell" us about the origin of language? In A. Vilain, J.-L. Schwartz, C. Abry, & J. Vauclair (Eds.), *Primate communication and human language: Vocalisation, gestures, imitation and deixis in humans and non-humans* (pp. 39–70). John Benjamins Publishing Company. https://doi.org/10.1075/ais.1.04lem

Liang, K.-Y., & Zeger, S. (1986). Longitudinal data analysis using generalized linear models. *Biometrika, 73*(1), 13-22.

Liszkowski, U., Brown, P., Callaghan, T., Takada, A., & De Vos, C. (2012). A prelinguistic gestural universal of human communication. *Cognitive Science, 36*(4), 698-713. http://doi.org/10.1111/j.1551-6709.2011.01228.x.

Locke, J. L. (2006). Parental selection of vocal behavior: Crying, cooing, babbling, and the evolution of language. *Human Nature, 17*, 155-168. https://doi.org/https://link.springer.com/article/10.1007/s12110-006-1015-x

Locke, J. L. (2009). Evolutionary developmental linguistics: Naturalization of the faculty of language. *Language Sciences*, *31*, 33–59.

Long, H. L., Bowman, D. D., Yoo, H., Burkhardt-Reed, M. M., Bene, E. R., & Oller, D. K. (2020). Social and endogenous infant vocalizations. *PloS One, 15*(8), e0224956. http://doi.org/10.1371/journal.pone.0224956

Lucca, K., & Wilbourn, M. P. (2018). Communicating to Learn: Infants' Pointing Gestures Result in Optimal Learning. *Child Development, 89*(3), 941-960. https://doi.org/https://doi.org/10.1111/cdev.12707

Lüke, C., Ritterfeld, U., Grimminger, A., Rohlfing, K. J., & Liszkowski, U. (2020). Integrated Communication System: Gesture and Language Acquisition in Typically Developing Children and Children With LD and DLD. *Frontiers in Psychology*, *11*, 118. https://doi.org/10.3389/fpsyg.2020.00118

Lynch, M. P., Oller, D. K., Steffens, M. L., Levine, S. L., Basinger, D. L., Umbel, V. (1995). Onset of speech-like vocalizations in infants with down syndrome. *American Journal on Mental Retardation, 100*(1), 68–86.

Masur, E. F. (1983). Gestural development, dual-directional signaling, and the transition to words. *Journal of Psycholinguistic Research*, *12*, 93–109.

McGillion, M., Herbert, J. S., Pine, J., Vihman, M., DePaolis, R., Keren-Portnoy, T., & Matthews, D. (2017). What paves the way to conventional language? The predictive value of babble, pointing, and socioeconomic status. *Child development, 88*(1), 156-166. https://doi.org/10.1111/cdev.12671

McNeill, D. (1992). *Hand and mind: What gestures reveal about thought*. University of Chicago Press.

Meier, R. P. (2006). The form of early signs: Explaining signing children's articulatory development. *Advances in the sign language development of deaf children*, 202-230.

Meier, R. P., & Willerman, R. (2013). Prelinguistic gesture in deaf and hearing infants. In *Language, gesture, and space* (pp. 401-420). Psychology Press.

Milenkovic P. (2010). TF32 [Computer Program]. *Department of Electrical and Computer Engineering, University of Wisconsin, Madison.*

Mitoyen, C., Quigley, C., & Fusani, L. (2019). Evolution and function of multimodal courtship displays. *Ethology, 125*(8), 503-515. https://doi.org/https://doi.org/10.1111/eth.12882

Moreno-Núñez, A., Rodríguez, C., & Miranda-Zapata, E. (2020). Getting away from the point: The emergence of ostensive gestures and their functions. *Journal of Child Language, 47*(3), 556–578.

Müller, G. B., & Newman, S. A. (2003). *Origination of organismal form: Beyond the gene in developmental and evolutionary biology*. MIT Press.

Mundy, P., Sigman, M., & Kasari, C. (1990). A longitudinal study of joint attention and Language development in autistic children. *Journal of Autism and Developmental Disorders*, *20*(1), 115-128.

Nathani, S., Ertmer, D., & Stark, R. (2006). Assessing vocal development in infants and toddlers. *Clinical Linguistics and Phonetics*, *20*(5), 351–369. http://doi.org/10.1080/02699200500211451

Newman, S. A. (2016). Origination, variation, and conservation of animal body plan development. *Cell Biology and Molecular Medicine Reviews, 2*, 130-162. https://doi.org/10.1002/3527600906.mcb.200400164.pub2

Nooteboom, S. G. (1999). Anatomy and timing of vocal learning in birds. In M. D. Hauser & M. Konishi (Eds.), *The design of animal communication* (pp. 63-110). MIT Press.

Oller, D. K. (1973). The effect of position-in-utterance on speech segment duration in English. *Journal of the Acoustical Society of America*, *54*, 1235-1247.

Oller, D. K. (1980). The emergence of the sounds of speech in infancy. In G.H. Yeni-Komshian, J. F. Kavanagh, & C. A. Ferguson (Eds.), *Child Phonology* (pp. 93–112). Academic Press. https://doi.org/10.1016/B978-0-12-770601-6.50011-5

Oller, D. K. (2000). *The emergence of the capacity for speech*. Lawrence Erlbaum Associates.

Oller, D. K., Niyogi, P., Gray, S., Richards, J. A., Gilkerson, J., Xu, D., . . . Warren, S. F. (2010). Automated Vocal Analysis of Naturalistic Recordings from Children with Autism, Language Delay and Typical Development. *Proceedings of the National Academy of Sciences*, *107*(30), 13354-13359. https://doi.org/https://doi.org/10.1073/pnas.1003882107

Oller, D. K., Buder, E.H., Ramsdell, H.L., Warlaumont, A.S., Chorna, L.B., & Bakeman, R. (2013). Functional flexibility of infant vocalization and the emergence of language. *Proceedings of the National Academy of Sciences, 110*(16),6318–6323. http://doi.org/10.1073/pnas.1300337110

Oller, D. K., Griebel, U., & Warlaumont, A. S. (2016). Vocal development as a guide to modeling the evolution of language *Topics in Cognitive Science (topiCS), Special Issue: New Frontiers in Language Evolution and Development, Editor, Wayne D. Gray, Special Issue Editors, D. Kimbrough Oller, Rick Dale, and Ulrike Griebel, 8*(2), 382-392.

Oller, D. K., Griebel, U., Iyer, S. N., Jhang, Y., Warlaumont, A. S., Dale, R., & Call, J. (2019a). Language origins viewed in spontaneous and interactive vocal rates of human and

bonobo infants. *Frontiers in Psychology*, *10*, 729. http://doi.org/10.3389/fpsyg.2019.00729

Oller, D. K., Caskey, M., Yoo, H., Bene, E. R., Jhang, Y., Lee, C. C., Bowman, D. D., Long, H. L., Buder, E. H., & Vohr, B. (2019b). Preterm and full term infant vocalization and the origin of language. *Scientific Reports*, *9*(1), 1-10. http://doi.org/10.1038/s41598-019-51352-0

Oller, D. K., Ramsay, G., Long, H. L., & Griebel, U. (2021). Protophones, the precursors to speech, dominate the human infant vocal landscape. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *376*(1836), 20200255. https://doi.org/doi:10.1098/rstb.2020.0255

Oller, D. K., & Griebel, U. (2021). Functionally flexible signaling and the origin of language. *Frontiers in Psychology*, *26*. https://doi.org/https://doi.org/10.3389/fpsyg.2020.626138

Orr, E. (2018). Beyond the pre-communicative medium: A cross-behavioral prospective study on the role of gesture in language and play development. *Infant Behavior and Development*, *52*, 66–75. http://doi.org/10.1016/j.infbeh.2018.05.007

Özçalışkan, Ş., & Goldin-Meadow, S. (2005). Gesture is at the cutting edge of early language development. *Cognition*, *96*(3), B101-B113. https://doi.org/https://doi.org/10.1016/j.cognition.2005.01.001

Özçalışkan, Ş., Adamson, L. B., Dimitrova, N., & Baumann, S. (2018). Do Parents Model Gestures Differently When Children's Gestures Differ? *Journal of Autism and Developmental Disorders*, *48*(5), 1492–1507. https://doi.org/10.1007/s10803-017-3411-y

Paulus, M., & Fikkert, P. (2014). Conflicting Social Cues: Fourteen- and 24-Month-Old Infants' Reliance on Gaze and Pointing Cues in Word Learning. *Journal of Cognition and Development*, *15*(1), 43-59. https://doi.org/10.1080/15248372.2012.698435

Petitto, L. A., & Marentette, P. F. (1991). Babbling in the manual mode: Evidence for the ontogeny of language. *Science*, *251*(5000), 1493-1496.

Pika, S., Liebal, K., Call, J., & Tomasello, M. (2005). Gestural communication of apes. *Gesture*, *5*(1-2), 41-56. https://doi.org/10.1075/gest.5.1.05pik

Pollick, A. S., & De Waal, F. B. (2007). Ape gestures and language evolution. *Proceedings of the National Academy of Sciences*, *104*(19), 8184–8189. http://doi.org/10.1073/pnas.0702624104

Prieur, J., Barbu, S., Blois-Heulin, C., & Lemasson, A. (2020). The origins of gestures and language: History, current advances and proposed theories. *Biological Reviews*, *95*(3), 531–554.

Ramos-Cabo, S., Vulchanov, V., & Vulchanova, M. (2019). Gesture and language trajectories in early development: An overview from the autism spectrum disorder perspective. *Frontiers in Psychology*, *10*, 1211.

Riede, T., Bronson, E., Hatzikirou, H., Zuberbühler, K. (2005) Vocal production mechanisms in a non-human primate: morphological data and a model. *Journal of Human Evolution, 48,* 85-96. http://doi.org/10.1016/j.jhevol.2004.10.002

Rivas, E. (2005). Recent use of signs by chimpanzees (pan troglodytes) in interactions with humans. *Journal of Comparative Psychology*, *119*(4), 404. https://doi.org/10.1037/0735-7036.119.4.404

Rodrigues, E. D., Santos, A. J., Veppo, F., Pereira, J., & Hobaiter, C. (2021). Connecting primate gesture to the evolutionary roots of language: A systematic review. *American Journal of Primatology*, *83*(9), e23313.

Roug, L., Landberg, I., & Lundberg, L.-J. (1989). Phonetic development in early infancy: A Study of four Swedish children during the first eighteen months of life*. *Journal of Child Language*, *16*(1), 19–40. https://doi.org/10.1017/S0305000900013416

Rowe, M. L., Özçalişkan, Ş., & Goldin-Meadow, S. (2008). Learning words by hand: Gesture's role in predicting vocabulary development. *First Language*, *28*(2), 182–199. https://doi.org/10.1177/0142723707088310

Silva Lima, E. D., & Cruz-Santos, A. (2012). Acquisition of gestures in prelinguistic communication: A theoretical approach. *Revista da Sociedade Brasileira de Fonoaudiologia*, *17*(4). https://doi.org/10.1590/S1516-80342012000400022

Sterelny, K. (2012). Language, gesture, skill: the co-evolutionary foundations of language. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1599), 2141–2151. https://doi.org/10.1098/rstb.2012.0116

Stark, R. E. (1980). Stages of speech development in the first year of life. In G. Yeni-Komshian, J. Kavanaugh, & C. Ferguson (Eds.), *Child Phonology* (p. 73-90). Academic Press.

Stark, R. (1981). Infant vocalization: A comprehensive view. *Infant Medical Health Journal, 2*(2), 118-128.

Stark, R. E., Bernstein, L. E., & Demorest, M. E. (1993). Vocal communication in the first 18 months of life. *Journal of Speech & Hearing Research*, *36*, 548-558.

ter Haar, S. M., Fernandez, A., A., Gratier, M., Knörnschild, M., Levelt, C., Moore, R. K., … Oller, D. K. (2021). Cross-species parallels in babbling: Animals and algorithms. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *376*(1836), 20200239. https://doi.org/doi:10.1098/rstb.2020.0239

Terrace, H. S., Bigelow, A. E., & Beebe, B. (2022). Intersubjectivity and the Emergence of Words [Review]. *Frontiers in Psychology*, *13*. https://doi.org/10.3389/fpsyg.2022.693139

Thal, D. J., & Tobias, S. (1992). Communicative gestures in children with delayed onset of oral expressive vocabulary. *Journal of Speech and Hearing Research, 35*, 1281–1289.

Tomasello, M., & Farrar, J. (1986). Joint attention and early language. *Child Development, 57*, 1454-1463.

Tomasello, M., & Zuberbühler, K. (2002). Primate vocal and gestural communication. In M. Bekoff, C. Allen, & G. M. Burghardt (Eds.), *The cognitive animal: Empirical and theoretical perspectives on animal cognition* (p. 293–299). MIT Press.

Tomasello, M., & Call, J. (2007). Ape gestures and the origins of language. In *The gestural communication of apes and monkeys* (pp. 221–239). Taylor & Francis Group/Lawrence Erlbaum Associates.

Tomasello, M., Carpenter, M., & Liszkowski, U. (2007). A new look at infant pointing. *Child Development*, *78*(3), 705–722. https://doi.org/10.1111/j.1467-8624.2007.01025.x

Tomasello, M. (2010). *Origins of human communication*. MIT press.

Trevarthen, C. (1979). Communication and cooperation in early infancy. A description of primary intersubjectivity. In M. Bullowa (Ed.), *Before speech: The beginnings of human communication* (pp. 321-347). Cambridge University Press.

Trevarthen, C. (2001). Infant Intersubjectivity: Research, Theory, and Clinical Applications. *Journal of Child Psychology and Psychiatry*(42), 3-48.

Volterra, V., Caselli, M. C., Capirci, O., & Pizzuto, E. (2005). Gesture and the emergence and development of language. In M. Tomasello & D. I. Slobin (Eds.), *Beyond nature-nurture: Essays in honor of Elizabeth Bates* (p. 3-40). Lawrence Erlbaum Associates.

West-Eberhard, M.J. (2003). *Developmental plasticity and evolution*. New York: Oxford University Press.

Wu, Z., & Gros-Louis, J. (2015). Caregivers provide more labeling responses to infants' pointing than to infants' object-directed vocalizations. *Journal of Child Language*, *42*(3), 538-561. https://doi.org/10.1017/S030500091400022

Yu, C., & Smith, L. B. (2012). Embodied attention and word learning by toddlers. *Cognition*, *125*(2), 244-262. https://doi.org/https://doi.org/10.1016/j.cognition.2012.06.016

Zuberbühler, K. (2017). The primate roots of human language. *Primate Hearing and Communication*, 175–200.

Chapter 2 Appendices (from Burkhardt-Reed et al., 2021)

Appendix A: Supplementary Background

Background on the Gestural Origin Theory of Language

Literature on the origins of language in recent decades has included an enormous amount of discussion of the idea that gesture was the primary mode of communication in our distant ancestors (e.g., Arbib, Liebal, & Pika, 2008; Hewes, 1973). As the reasoning goes, gesture was elaborated into primitive language (some sort of protolanguage) in an early phase of evolution away from the pan-primate background (Sterelny, 2012; Tomasello, 1996, 2008), and vocalization had to have taken over as the primary mode of communication in recent time, co-opting somehow the roles that had been primarily gestural previously (Armstrong, 2008; Corballis, 2002). Further, in accord with the reasoning, the shift to vocal language left patterns where gesture plays very salient, even sometimes purportedly predominant roles, in modern language learning (Iverson, 2010; Iverson & Fagan, 2004). The reasoning associated with this gestural origins theory has been repeated enough times that it has become for practical purposes a sort of standard wisdom about the origin of language in much of the relevant literature (Irvine, 2016), in spite of substantial objections that can easily be enumerated (e.g., Masataka, 2008; Seyfarth, 2005). A search of Google Scholar on "gestural origin language" returned thousands of citations.

The final author of this paper has long argued that the gestural origins theory has been overplayed (Oller, 2000), substantially based on his experience in longitudinal research in vocal development starting as far back as the 1970's (Oller, 1978, 1980; Oller, Wieman, Doyle, & Ross, 1976), where signs of early gestural communication have always seemed minor by comparison with the vast evidence of vocal activities of infants. The viewpoint has never

rejected a role for gesture in language evolution and development, but has been inclined to support a primarily vocal origin, supplemented by significant, but more limited gestural features. Recently the Origin of Language Laboratory (OLL) at the University of Memphis, directed by Oller, have been in pursuit of direct comparisons of gesture and vocalization as early communication vehicles.

Background on Evolutionary Developmental Biology (Evo-devo) and Its Implications for the Gestural Origins Theory

Evo-devo is a central theme of modern theoretical biology (Newman, 2016; Newman & Müller, 2000). Evo-devo reasoning regarding the origin of language has been addressed in a variety of publications from OLL projects (Griebel & Oller, 2008; Griebel, Pepperberg, & Oller, 2016; Long et al., 2020; Oller et al., 2013; Oller & Griebel, 2014, 2021, 2008; Oller, Griebel, et al., 2019; Yoo et al., 2018). A basic tenet of evo-devo is that natural selection does not build whole new structures or capabilities in mature organisms—rather natural selection targets developmental processes (Carroll, 2005; West-Eberhard, 2003), such that new and more elaborate mature structures or capabilities can emerge from changes in early stages of development. These relatively small changes in individuals can be augmented cyclically by natural selection across generations if the conditions of selection are right, with development reflecting selection targets in each generation.

Because this is the way evolution typically operates, early developmental patterns can be expected to reflect the natural logic of both the development and the evolution of any system. Consistent with evo-devo logic, one would predict, if gesture forms the primary basis for human communication, gesture should predominate in the earliest communications of humans. The fact that gestures were relatively infrequent in our data, and furthermore that they were less likely to

be socially-directed than protophones (even though receiver gaze is *required* for gesture to be effective communicatively), suggests the reasoning underlying the gestural origins hypothesis is flawed based on the perspective of modern theoretical biology.

On the High Rates of Protophone Production in Human Infants

Until relatively recently it was possible to assess vocalization frequency in human infants *only* based on laboratory recordings or short-term recordings conducted in homes. This left it uncertain whether vocalization and interaction frequencies occurring in available data were representative of the frequencies occurring in daily life. But starting in ~2008, it has been possible to make estimates based on all-day audio recordings in infant homes. Based on both automated analysis of these recordings (e.g., Gilkerson et al., 2017), and human coding of randomly-selected samples from them (e.g., Oller, Caskey, et al., 2019), it is now known that human infants produce massive amounts of protophone vocalization. Even when they are alone and awake in a room, the rates are 3-5 per minute. In laboratory settings the rates appear to be perhaps twice as high, but it is uncertain why. It could be that in laboratory recordings infants are maximally awake and alert, rarely drowsy, while random sampling of presumed wakeful segments from all-day recordings may include many segments where infants are not asleep, but are nonetheless drowsy. In OLL laboratory recordings, parents are asked to bring infants in during periods of likely full wakefulness, and a drowsy infant may prompt us to terminate the recording and reschedule.

Perhaps surprisingly, in relatively controlled laboratory recordings, infant protophone rates are not higher during face-to-face interaction segments than during segments where the parent is silently occupied, for example, reading a magazine (Iyer, Denson, Lazar, & Oller, 2016). Other vocal types of interest are human infant cry and laughter, both of which are salient when they

occur—but protophones occur at least 5 times more frequently across the entire first year, and are even much more frequent than cry in premature infants still in neonatal intensive care (Oller, Caskey, et al., 2019). Moreover, human infant protophone rates are massively higher than rates of protophone-like sounds occurring in bonobo and chimpanzee infants, at least 10 times higher (Oller, Griebel, et al., 2019).

It seems likely that the false impressions of many that crying is more frequent than protophones and that infant face-to-face interaction generates most protophones are the result of selective attention. We tend to remember that which we notice and attend to. Crying and face-to-face interaction may produce more salient memories than protophones produced in infant comfort while caregivers are otherwise engaged.

On Comparing Gesture and Vocalization—the Apples-to-Oranges Problem

One might imagine that comparison of the amount (frequency) of vocalization vs. gesture is comparing "apples to oranges". Of course the modalities are different, a fact which presents a comparative challenge, but that challenge has been with us for as long as there has existed a gestural origins theory (from at least as far back as Hewes, 1973). One has to address the problem head on in order to make a useful comparison, and that has been one of the primary goals of our recent efforts. We have recognized that a solid foundation for the method of comparison has been needed, and our infrastructural approach to establishing reasonable units of comparison across the modalities represents, we think, an advancement.

The coding approach we have formulated targets the *functions* that can be served communicatively by either modality so that a comparison can be justified and meaningful, at least at that functional level of abstraction, a comparison of apples to apples. Both gestures and vocal events can serve as communications both in prelinguistic phases or in conventional

communications as signs or words. If we can compare the numbers of signs to words in the emerging lexicon of, for example, a child growing up in a home where both sign language and spoken language are active, surely, we reason, one can also compare the numbers of precursors to signs to the numbers of precursors to words.

Our coding scheme is designed precisely to focus on the event-level communicative functions or potential communicative functions of developing actions in the gestural and vocal domains. Each such event can be counted: an infant hand position, the d-hand for example, is a gestural component that might be incorporated into a sign, and a normal phonation event is a vocal component that might be incorporated into a word. Sometimes these actions actually function communicatively for the prelinguistic infant, but even when they do not, we argue that they provide a foundation for possible communication in the future. Thus, the relative numbers of events produced in the two modalities across time provide a measure of the relative roles of the two modalities in communication or potential communication.

However, the apples-to-oranges problem is inherent: differences in the modalities create a challenge. The nut of the matter is that the visual/gestural modality sometimes exploits opportunities that the auditory/vocal modality cannot exploit, and vice versa. To realistically make quantitative comparisons across modalities, these differences must be recognized. We made a specific effort in the current research (because we are aware of the popularity of the gestural origins account of the origin of language) to make these comparisons in such a way that, in cases of ambiguity about how counts should be done, our procedure would be biased to ensure inclusion of gestural material more than vocal material. For explanation on this intentional bias, see below under **Supplementary Methods**, **Ensuring that Gestures Were not Under-counted Compared with Protophones.**

Consider the apples-to-oranges problem in terms of differences in opportunities for communication in the visual/gestural and auditory/vocal modalities. The idea of "fixed signals" in classical ethology (Lorenz, 1951; Tinbergen, 1951) provides perspective; the visual/gestural modality is naturally capable of conveniently and efficiently transmitting information about location in space using physical movements, a type of fixed signal "deixis" that cannot be replicated in the vocal domain. Pointing, for example, can be viewed as a fixed signal because it is inherently deictic—it does not have and cannot acquire other functions in infancy. In the vocal domain, deixis can involve special lexical items such as 'this', 'that' and 'there', all of which depend on lexical syntactic context to function as "pointers" independent of gesture. These "words" are not fixed signals, because they have to be learned.

So, gesture has a special role in providing efficient here-and-now deixis with fixed signals such as pointing, showing, and reaching toward an object. Vocalization, on the other hand, is naturally capable of conveniently and efficiently transmitting here-and-now information about affect, as in crying, whimpering or laughing, all of which are fixed signals of the vocal domain of human infancy; they, like the deictic signals of gesture, have single functions that cannot be fundamentally modified. The vocal domain also supplies particularly effective transmission of *negative* affect in whining during talking or protophone production (Jhang, Franklin, Ramsdell, & Oller, 2017). Gesture on the other hand has more limited potentials for transmitting affect, independent of facial or vocal accompaniments. These differences highlight the apples-to-oranges problem for our enterprise.

Notice that the literature on the gestural origin of language has routinely taken pointing and other physical deixis actions as being inherently a part of language, while ruling cry and laughter out of consideration as foundations for language. This pattern in the literature, we think,

represents an important bias, which has contributed to obscuring the actual role of pointing and other physical deixis actions in language. Such actions are confined to the here and now, and experience cannot fundamentally change the functions of the deictic gestural actions. Other gestural actions (such as special hand positions or arbitrary but systematic movements of hands or arms) have the potential to function as learned lexical-like material, arbitrarily associated with semantic content. Such lexical material, in signs, can be detached from the here and now to transmit information about events elsewhere in the world, about the past, the future or the imagination. Such gestural actions are thus not fixed signals. In the same way, vocal material, such as particular syllables, can be arbitrarily associated with semantic content in words, which can be detached from the here and now to transmit information outside the present. Our counts of gestures and protophones in the present study were designed to quantify the events that could serve as "specialized" communications (in the sense of Hockett (Hockett, 1977; Hockett & Altmann, 1968)) in both domains, but we have systematically biased the counts to favor gestures.

Even with this biasing, protophones were more frequent, especially at the youngest age, where gesture was nearly absent (a grand total of 46 gestures were observed, 33 of which were produced by 2 of the 10 infants), but protophones were plentiful (with almost 1700 protophones having been documented, and the infant with the smallest number at that age produced 85 protophones). It is important not only that protophones occurred far more frequently than gestures in our data, but also that the proportion of cases where a *protophone* was directed by gaze to another person was twice as high as the proportion of cases where a *gesture* was directed by gaze to another person—this finding suggests that to the extent that either modality is being developed for communication, *gesture is clearly not playing the larger role*. In other words, we

think, the data supply substantive information to doubt a primary role for gesture in the origin of language.

Our point is also buttressed by the fact that vocalizations can be heard independent of gaze, and consequently if an individual wishes to communicate by gesture, the importance of ensuring that the receiver is looking is surely higher in the case of gesture than of vocalization, and yet vocalizations showed much higher proportions of gaze toward the potential listener than gestures did. These quantitative facts are also independent of the fact that in both modalities, during the first year, most of the relevant acts are not directed to anybody (Long et al., 2020). Having located the events that were socially directed (using gaze as the indicator), the data show that events of vocalization were far more directed than events of gesture.

The Importance of Comparing Frequency of Occurrence of Protophones and Gestures in Early Human Development

The claims about the gestural origin of language represent substantive assertions about the infrastructure of language, its evolution, and its development in modern human infants and young children. In this form the theory has often involved empirical evidence, presented in quantitative form, purported to show that early vocabulary in children is in fact heavily gestural, and that early sentences are heavily composed of gestures and words, and often of gestures alone (Titze & Strong, 1975; Tomasello, 1996, 2008). So there has routinely existed an assumption in the discussion of the role of gesture in both language development and language evolution that frequency of occurrence of gestures and verbalizations form evidence regarding the relative importance of each.

Gesture, vocalization, and facial affect all form important aspects of human communication, and all three are involved in the first year. Furthermore, all three can be concluded to have formed

important aspects of ancient hominin communication, just as all three play roles in the communication systems of our primate relatives. It is important, however, to differentiate among the aspects of communication that rely on each of these modalities and to *quantify* their relative roles.

No one seems to doubt that the information transmitted in language is primarily vocal in *mature* humans (Irvine, 2016) (with no hearing disorder). This conclusion is surely based on the apparent fact that the vast majority of lexical and syntactic material is transmitted vocally by humans (or in a written form derived from the vocal one). Key claims of gestural origins advocates are that quantitatively, gesture plays a primary role in early language, that gesture plays a quantitatively more important role than vocalization in communication in all the great apes other than humans, and that consequently ancient hominins likely used quantitatively more gesture than vocalization in communication—and most importantly that the earliest language-like behaviors of hominins were gestural (Tomasello, 2008). Even though some of the gestural-origins literature has argued from the arm-chair, without quantifying its claims, much of the literature has been thoroughly quantitative, arguing on the basis of counts of communicative events that gesture is more important in very early human development than vocalization and more important in non-human primate communication than vocalization. So, the foundation for making quantitative comparisons of amounts of gesture and vocalization in the context of theorizing about the origin of language is deeply established.

Moreover, frequency of occurrence is treated seriously in every domain of action under assessment for development, as far as we can tell. Scientists conclude that a child has developed a consolidated command of walking when the child takes steps regularly and doesn't fall or need to hold on to a table, not when there is a single step or a few steps taken occasionally.

Development of walking is assessed by frequency of occurrence of walking acts or of individual steps evaluated over time (Adolph, Robinson, Young, & Gill-Alvarez, 2008). Scientists assess early vocabulary development in terms of how many words a child uses or understands, how often they use them, and how long the utterances are that children use at various points, all of these being matters of frequency of occurrence (Brown, 1973; Fenson et al., 2007). The same is true of evaluation of sign language acquisition.

It is now known that the human infant produces on average ~3500 protophones daily while bonobo and chimpanzee infants produce protophone-like sounds at a rate not even 1 tenth as large (Oller, Griebel, et al., 2019), a frequency of occurrence comparison that reveals a likely selection pressure on vocal action in the hominin line; we emphasize that this fact could easily go unrecognized without the quantitative comparison. Note that a non-quantitative binary judgment (very common in comparative biology) might claim that humans are like bonobos and chimpanzees because both humans and the ape cousins have protophone-like sounds. This sort of non-quantitative, binary judgment has fostered a confusing literature in animal behavior that has tried to support the claim that all features of language are present in communications of animals, obscuring and failing to illuminate the differences between human and animal communication that everyone seems to acknowledge (Snowdon, 2004). This confusion is a direct result of the failure to quantify comparisons.

The Importance of Pointing and Other Forms of Gestural Deixis in the Origin of Language

Our results quantitatively challenge the widespread belief in a primarily gestural origin of language, but at the same time they do not deny or contradict the importance of gestural phenomena in helping to bootstrap vocabulary learning and to support development of

communication in general. That pointing is foundational is acknowledged extensively in prior work from the OLL (especially Oller, 2000; Oller, Griebel, & Warlaumont, 2016).

In fact, there is a particular kind of function that pointing (and other deictic gestures such as reaching toward an object or holding one up to show it) can perform that is not possible for a preverbal infant to perform vocally. Pointing and other deictic gestures can designate an entity for joint attention and thus make it possible for labeling to be at least partially disambiguated. Deictic gestures continue to play important roles in mature communication as well, in part because they are sometimes more efficient than verbal deixis. For example, if one points to a novel object, designating it, one might understand that attention is being drawn to the novel object, and if the attention is indeed shared by another person, a name for the object might be given by the first party and learned by the second. Second language learning often proceeds this way to some extent. Through verbalization alone, the same process would be considerably more complicated. One might say "I am attending to a novel object on the table in front of us"—that's quite a lot of verbiage to designate the object compared to the simple pointing gesture. Still, language includes enormously more than deixis

Pointing and Other Forms of Gestural Deixis are not Symbols

A key issue is that pointing is not (nor is any other form of gestural deixis), in and of itself, language, because language at its very foundation requires symbolism, and linguistic communication is overwhelming symbolic (de Saussure, 1968; Deacon, 1997; Peirce, 1934; Sinha, 2004). Labeling by words or signs is a kind of symbolism in this sense. Pointing is a kind of visual deixis designating entities that *can be* labeled (that is, referred to by symbols), but pointing cannot itself label and is thus not symbolic—it is not a vocabulary item that can make reference to an entity at any time, including any time outside the here and now. The great

advantage of pointing (and other forms of gestural deixis) is that it can *scaffold the development of labeling* (either in the spoken language or the sign language domain) using an action which the vocal modality cannot replicate, but can only perform with words or sentences, and never so efficiently during early development as in the case of the gestural acts.

Yet our data showed pointing by infants occurred very infrequently in the interactions we observed in the first year of life—fewer than 2% of the gestures we observed were points, all of them at the latest age, and all of them from two of the ten infants. It seems likely that once word learning is well underway in the second year, pointing may play a much more prominent role and occur much more frequently. But, even at later ages, pointing remains a kind of visual deixis, and cannot become an element of the lexicon.

Directivity of Communication or Potential Communication, and Relative Lack of It, in Human Infancy

One might ask why most gestural and vocal acts of potential communication are not directed to anyone. This is an important question to which we have supplied extensive response (Oller & Griebel, 2021; Oller et al., 2020). The key idea is that vocal exploration (the production of protophones) has been naturally selected as a fitness (or wellness) signal. Hominin infants have long been altricial (helpless) and in need of especially high investment from caregivers for several years of growth during which they cannot forage for (or protect) themselves. As the reasoning goes, hominin infants, because of their altriciality, have long been under pressure to supply indicators of their wellness, indicators that would influence caregivers to keep them and care for them even in times of hardship (Locke, 2006, 2009; Oller & Griebel, 2005). Vocalization, according to the reasoning, came to provide a means of fitness signaling in response to the pressure imposed by altriciality. Gestures could have similarly come under

selection pressure. However, vocalizations can be heard even if caregivers are otherwise occupied, not looking at infants. Consequently, *protophones don't need to be directed* (although sometimes they are) in order to function as fitness signals as long as caregivers hear them. Gestures, on the other hand, need to be seen in order to function as fitness signals, and consequently there is good reason to suppose that selection pressure on gestures as fitness signals would have been lower than on vocal signals.

The fact that most infant vocalizations and gestures are not socially-directed suggests that they are largely endogenously generated. It can be reasoned that human infants are endowed with *an exploratory motivation* that includes vocalization and gesture, a motivation that presumably serves them well as practice that may be useful in later language learning at the same time that it serves to signal their well-being (Locke, 2006; Oller & Griebel, 2005). Also, the fact that infants do not direct their vocalizations to caregivers most of the time does not imply that caregivers do not notice them and form opinions about the well-being of infants on the basis of noticing them. In fact it is the caregivers' actions in response to vocalization of infants that is argued to be the primary basis upon which the endogenous vocalization tendency of infants has been naturally selected (Locke, 2006). And notice that the endogenous vocal tendency continues throughout human life, with adults producing huge amounts of vocalization (Mehl et al., 2001; 2007), much of it playful, not particularly informative, and often not even directed to another person.

A comprehensive theory of language development has to account for two central facts that may at first blush seem contradictory: infants explore vocalization (and gesture) and presumably develop vocal (and gestural) categories *endogenously*, and at the same time infants learn about the utilization of vocalization (and gesture), and ultimately about the relevant categories for their ambient language, through *listening and vocal interaction*. We do not view these facts as being

contradictory, but rather as two features of evolved human inclination, both of which contribute (presumably critically) to the development of the language capacity. These seemingly contradictory tendencies do not, in our view, oppose, but rather supplement each other.

The Importance of Dyadic Communication in Infancy

Our goal in this research was not primarily to monitor joint attention or triadic communication, but rather to monitor all communicative acts (the great bulk of which at these ages are dyadic or monadic, not triadic) including the efforts of infants to direct their potentially communicative acts toward a caregiver. Some child developmentalists have tended to neglect non-triadic communications, not treating them as communications at all. We take a different viewpoint, accepting acts of purely dyadic interchange as among the most important of communicative events for infants and their caregivers. For example, if an infant simply alternates vocalizations with a parent face to face, we deem this interaction to be communicative. Similarly, an expression of distress directed to a parent, even if it does not indicate what the distress is about, is important communication in our view. Even an expression of distress that is not intended as a communication by an infant may be interpreted as one by a caregiver. So, there is a great deal of communication that goes on in early human infancy.

We also have argued extensively that extensive dyadic communication (the primary intersubjectivity of Trevarthen (Trevarthen, 1979, 2001)), especially face-to-face vocal interaction, is a foundation that makes triadic communication possible. The idea is that one cannot engage in joint attention to an entity separate from the two parties to an interaction unless one understands that the interlocutor has the possibility of sharing awareness or knowledge with another. Thus, we argue, in agreement with Trevarthen, that dyadic communication needs to be deeply developed in order to make triadic communication possible, because repeated dyadic

communication instills in the infant an awareness of the possibility of sharing awareness or knowledge with another. In accord with this reasoning, it should be no surprise that pointing and other acts of joint attention develop later than extensive face-to-face interactions.

Triadic communication, such as pointing, was very infrequent in our data. Acts of showing (which we also interpret as triadic), were even less frequent. We are not the only researchers to have reported low rates of pointing in interaction between 12-month olds and parents (6-8 per 30 min (Gros-Louis & Wu, 2012)). The pointing rates seem low especially when they are compared to protophone rates (4-8 per min depending on circumstances (Oller, Griebel, et al., 2019)).

Supplementary Methods

Recording Details

The recording circumstances for the present paper have been described in considerable detail in prior publications from the OLL (Oller et al., 2013; Oller, Caskey, et al., 2019; Oller, Griebel, et al., 2019) with amplifications in the Supplementary Materials to those papers. Here we offer additional details.

This was a naturalistic study conducted in quiet spaces designed to resemble a child's playroom with an adjacent control suite to allow viewing through a one-way mirror as well as including monitors for all the cameras operative in the recording room. Because we sought naturalistic data, we allowed the parents, who were always present during the segments of recording used for this study, to move about freely with their infants. Thus, there were no restrictions on movement during the recordings except those that may have been imposed by the parent on the infant. The recordings were intended to maximize visibility (multiple cameras, with two recording at each point in time) and audibility (microphones worn on the infants and parents, with separate channels) of the participants.

The periods we evaluated in this study were all designated as "interaction periods" for the parents. Of course, infants were more mobile at the later ages, when, we suspect vocalization rates were perhaps somewhat depressed (see Figure 2, main text) because of high mobility and infant distraction from vocal interests by the things they found to do in the room. There were always toys available, generally relatively silent ones, and never in large numbers, since we did not wish to have the infants so engaged in toys that they would not interact communicatively. However, the latest age did not correspond to lower *gesture* rates. A major difference across ages was that the earliest age (when the infants were not even able to sit up unsupported) produced very low gesture rates (about 7 times lower than at the later ages), but very high vocal rates. It remains possible that the relatively lower rates of protophone production at the middle and late ages were simply a sampling artifact since other research has not clearly supported a reduction in protophone rates across the first year (Gilkerson et al., 2017; Iyer et al., 2016).

The parent was always in the same room and visible on camera. Of the two cameras that were recording at each point, our protocol specified that one camera was to be focused on the infant (especially the face and torso) to the extent possible, while the other was focused on "the interaction" (intended to include both the infant and any other participants in the interaction). Since cameras were available in all four corners of the room, the person managing the cameras usually was able to follow the protocol with no difficulty, switching cameras with changes in positioning of the participants.

On Methods for Monitoring Vocalization and Gesture in the First Year of Life

The methodological problems of monitoring gesture and vocalization in early life have a good bit in common, as we see it, having worked on the comparison extensively over the past few years. We have written a great deal about optimal methods and difficulties of monitoring vocalization (see Oller, 2000, and Supplementary Material to Oller et al., 2013; Oller, Caskey, et al., 2019; Oller, Griebel, et al., 2019). Most of the same difficult features of monitoring apply to gesture coding, and to a similar extent.

A special difficulty of monitoring gesture during development is that the gestural organs (especially the hands and arms) have many non-communicative (we call them "utilitarian") functions (such as touching things to palpate them, picking things up or eating) that confuse their possible roles as communicative vehicles. In the vocal domain there are also non-communicative, utilitarian functions (these are inherent in vegetative sounds such as coughing, sneezing, burping, …) that need to be interpreted with caution. Our approach to comparison has been to leave out purely or primarily utilitarian acts of both domains, but we are open to suggestion from others who may take up the challenge of addressing the gestural origin theory quantitatively.

One might suggest instrumental motion tracking, rather than human coding, as a means of monitoring gesture and vocalization. In instrumental monitoring, similar issues apply to both gesture and vocalization. Human speech has been argued to be the most complex motor act we perform (e.g., Simonyan, Ackermann, Chang, & Greenlee, 2016), and there is a vast literature on direct monitoring of the movements of the supraglottal tract in speech, as well as the motoric control of the glottis, along with the acoustic consequences of vocalization, research from the field of "speech science" (Borden, Harris, & Raphael, 2011; Kelso, Saltzman, & Tuller, 1986;

Kent, 1998). While motion tracking or acoustic analysis can supplement, they cannot substitute for human judgments about the nature of either gesture or vocalization as communications (see discussion in Oller, 2000). The human observer is the source of selection pressure on any form of human communication, and any method of assessing such communication is ultimately bound to the human observer as the ultimate relevant source of judgment. For both vocalization and gesture, if it were not possible to perceive the actions and to interpret them appropriately (as communications), it would be impossible for a communication system (either vocal or gestural) to exist in any species (Maynard Smith & Harper, 2003)—without consistent perceptual interpretation, the potential communications would disintegrate across evolution for failure to be understood. So, it is our opinion that in either the gestural or the vocal domain, the primary judgments must be human-perception based, and motion monitoring or acoustic analysis must be viewed as secondary approaches. In the OLL projects, we use acoustic analysis extensively (Buder, Chorna, Oller, & Robinson, 2008; Buder & Oller, 2019), but always maintain human perception as the primary target of evaluation.

Motor development across the first year involves both the vocal system and the gestural system. It appears that there exists very early motoric development in the vocal domain to a greater extent than in the gestural—the data from the present work do indeed imply that motoric development of some key features of vocalization precede the development of key features of gesture. Why might that be so? A distinct possibility, as we see it, is that selection pressure through hominin history has favored early development of vocal capabilities precisely because they were more important in communication of young hominins than gesture was (or other kinds of motoric development were). So, the apparently relatively late development of some gestural capabilities can be viewed as evidence of its own against the gestural origins theory**.**

Ensuring that Gestures Were not Under-counted Compared with Protophones

We implemented specific procedures for the present study, designed to avoid the possibility of overplaying the amount of vocal communication. So, we biased our counts to favor gestures in each of the following cases.

1) Universal social gestures, such as pointing, showing, or reaching for an object that cannot be attained by an infant independently, all have clearly definable communicative (or potentially communicative) deictic functions that reveal them to be, in the sense of classical ethology, "fixed signals" (Lorenz, 1951; Tinbergen, 1951)—the gestural signals are not arbitrarily associated with their functions, but are instead fixed by nature to those functions. We always included those gestural "fixed signals" in our counts. By the same token crying, whimpering, and laughter are fixed signals of the vocal domain, having fixed functions, but we never included them in our counts. Had we included them, prior results indicate they would have increased the vocal counts by at least 10-15% (Oller et al., 2013).

2) We never included vocal vegetative sounds (coughs, sneezes, effort grunts…) in our counts, even though there are occasions when vegetative sounds yield communicative import, because parents take action to assist in response to them, and sometimes infants (even as early as the middle age in our study) produced them as playful communications. Normally one might call vegetative sounds "utilitarian" in a similar sense to the way we used the term "utilitarian" for physical movements that we did not include in our gesture counts, such as simply touching an object or grasping and picking it up. Yet we always included "reaches" as gestures, as long as the object reached for was not actually successfully reached by the infant independently of help from the caregiver. In both the physical movement and vocal cases, "utilitarian" actions could be thought of as naturally directed to a goal independent of communication, but we included in our

gestural counts some cases of physical movements (reaches) that may have had no communicative import, while we ruled out all cases of vegetative sounds regardless of their potential to communicate.

3) Vocalizations included considerable material that was always excluded even though one might argue it should have been included. We counted only the vocalizations that could be gauged to constitute vocants (vowel-like sounds with clearly audible phonation), squeals (high pitched phonated sounds) or growls (low-pitched or harshly phonated sounds)—these three included utterances with both canonical and non-canonical syllables, the canonical squeals, growls and vocants occurring primarily at the late age, although non-canonical protophones were predominant at all three ages.

Importantly, however, the utterances we counted did *not* include:

a) non-phonated sounds even if they were clearly communicative, such as whispers,

b) voiceless frication sounds or bursts,

c) fully audible ingressive vocalizations (often with phonation) that sometimes occurred independently but often also occurred in between utterances, or

d) any sound (even if it would otherwise have been coded as vocant, squeal or growl) that the coder deemed to be so low in intensity or so short in duration, that it would likely not have been attended to by a caregiver.

We implemented these procedures for the present study, biasing our counts to favor gestures in each case, in order to produce a body of information that would be hard to reject as being unfair to the gestural origins theory.

We also selected the interactive sessions rather than sessions where parents were engaged in reading (no-talk-to-baby sessions) or interview with another adult on the assumption that

gestures would more likely occur in the interactive sessions, since parents were much more likely to look at the infants during those sessions than during the no-talk-to-baby or interview sessions. Since we intended to assess the gestural origins theory, this was another way we sought to avoid an approach that might bias our observations in such a way as to limit the occurrence of gestures.

How to Gauge Directivity of Communication and Potential Communication in Human Infancy

Even though most vocal and gestural actions of infants tend *not* to be directed socially, scientific interest in vocal and gestural development is largely driven by the role of these modalities in communication, which in its prototypical forms is assumed to be socially directed. So, we seek to portray directivity, because it suggests intentional communicativeness, presumably growing across time in the infant.

We sought a relatively simple way to gauge social directivity, and settled on gaze direction as the best available measure that could be obtained with relative coding ease. Clearly other factors also play roles in natural judgments of social directivity (timing relative to utterances of other speakers, prior situational context, etc.), but there is good reason to use gaze as a particularly potent gauge because it correlates very highly with judgments of directivity based on other factors available to observers judging directivity based on audio-video recorded interactions, as demonstrated in an investigation conducted in the OLL (Long et al., 2020). Gaze direction as the primary directivity measure has also been used by others who have sought to code intentionality of gesture and vocalization in early childhood (Bates, 1976; Franco & Butterworth, 1996; Iverson, 2010; Iverson, Tencer, Lany, & Goldin-Meadow, 2000; Thal & Tobias, 1992).

Any instance of gaze toward a person during a gesture or during a vocalization (plus a 50 ms addition at the beginning and end of each event) was counted as an instance of social directivity in our research. So, the gaze toward a person did not have to coincide entirely with the gestural or vocal event, but only for a moment noticed by the coder. It would be of interest, of course, to code within say five seconds before and after the events to see if gaze was directed to someone during those periods. This is a task for future OLL research.

Supplementary Results

Details on Statistical Analyses

The primary patterns of interest in this research were unambiguous and could be confirmed with a variety of statistical methods. We did indeed conduct several tests for both of the main hypotheses, in addition to those reported in the main text, and the results confirmed rejection of the gestural origins predictions in all cases. These methods included a repeated measures ANOVA and a variety of t-tests for individual comparisons. But because of non-normality of distributions, these tests are considered secondary. We also conducted a series of (non-parametric) chi-square tests, results of which conformed to those reported in the main text based on GEE and followup Mann Whitney U tests.

The following are the Tables representing the GEE results reported in the main text in detail. In each case the Estimates represent the GEE estimated means. The standard errors of those means are also provided. The means and standard errors supply GEE effect size indicators. For example, for Table SM1 the Estimate of 165.1 with Std Err of 23.1 suggests an effect size of 165 more protophones than gestures.

Table 1. Hypothesis 1, GEE Analysis on Protophone and Gesture Rates (Age as Factor)
This table displays the GEE analysis on Protophone and Gesture rates with Age as a factor. Early Age is the baseline for Age, and for Modality, Gesture is the baseline.

Coefficient	Estimate	*SE*	*p*-value
Intercept	4.600	2.031	0.0235
Middle Age	28.00	5.579	5.19e-7
Late Age	33.3	6.247	9.78e-8
Modality	165.1	23.117	9.2e-13
Middle Age*Modality	-74.00	31.516	0.0189
Late Age*Modality	-106.10	29.535	0.0004

Table 2. Hypothesis 1, GEE Analysis on Protophone and Gesture Rates (Age as Continuous Variable)
This table displays the GEE analysis on Protophone and Gesture rates with Age as a continuous variable. For Modality, Gesture is the baseline.

Coefficient	Estimate	*SE*	*p*-value
intercept	-2.53	3.33	0.4473
Age	3.76	0.76	7.7e-7
Modality	188.04	35.69	1.4e-7
Age*Modality	-11.31	4.02	.0049

Table 3. Hypothesis 2, GEE Analysis on Directivity of Gestures and Protophones (Age as Factor)
This table displays directivity of gestures and protophones for the Age as factors model, where Middle Age is the baseline. For Modality, Gesture is the baseline.

Coefficient	Estimate	*SE*	*p*-value
Intercept	0.0619	0.0393	0.11518
Middle Age	0.0951	0.0693	0.16993
Late Age	0.1420	0.0671	0.03436
Modality	0.3429	0.0687	6e-7
Middle Age*Modality	-0.0917	0.0821	0.2642
Late Age*Modality	-0.2792	0.0812	0.00059

Table 4. Hypothesis 2, GEE Analysis on Directivity of Gestures and Protophones (Gesture as Baseline)
This table displays the GEE analysis on directivity of Gestures and Protophones for Age as a continuous variable. For Modality, Gesture as the baseline.

Coefficient	Estimate	*SE*	*p*-value
Intercept	-0.00157	0.05639	0.978
Age	0.01943	0.00774	0.012
Modality	0.51565	0.08125	2.2e-10
Age*Modality	-0.04041	0.00910	9.0e-6

Duration of Gestures and Protophones

The gestures were longer, on average about twice as long, as vocalizations in the data presented in the main text. Even with that greater duration per event, vocalizations still accounted for more time during the recordings than gestures: for the earliest age there was 17 times more vocalization time than gesture time (a fact that is mainly due to the very small number of gestures that occurred at the earliest age), while at the older ages the difference favored the vocalizations by ~75%.

For the purposes of evaluating the gestural origins theory, the event-level analysis is the appropriate one, we believe, for the primary comparison. Each event is a potential communicative act, and it is the number of such acts that is at stake in the determination of how often communications are transmitted in the two modalities. Consequently, we contend the event-based analysis is preferable for primary focus with regard to the gestural origins theory. In addition, the fact that the durations of gestures were greater (by a factor of about 2) than vocalizations, along with the fact that gestures were *less likely* than vocalizations to show directed gaze during any portion of those longer durations, adds yet more weight to the conclusion that the gestures were less likely intended as communications than the

vocalizations—remember that an unseen gesture has no communicative import, while vocalization can indeed be communicative without gaze-sharing.

Another issue to consider is that shorter duration of protophones might be taken to imply greater efficiency of communicative acts in the vocal domain, at least during infancy. Indeed, the vocal modality may have been selected as the primary form of early language in part because of its supreme efficiency. At the same time, we recognize that natural sign languages as used by mature native signers have been shown to be comparably efficient to spoken languages in the transmission of information (see Bellugi & Fischer, 1972 or various articles in the journal *Sign Language Studies*).

Facial Affect in Gestures and Protophones

Facial affect is a third modality of communicative expression in humans adding to vocalization and gesture. Throughout life, facial expressions accompany both spoken and signed languages, and are seen in precursors to both spoken and sign communication in early development and in maturity. Although facial affect often accompanies other modalities of communication, it is often also independently functional, occurring in the absence of any vocalization or gesture. Like vocalization, facial affect is displayed plentifully from the first day of life, while gesture, based on the present data, appears to begin its occurrence several months later.

We coded facial affect (positive, negative, neutral, can't see) according to the procedures described in Oller et al. (2013), during the period of each protophone and each gesture, adding 50 ms to the beginning and end of each period to ensure that coders could view the entirety of each gesture or protophone while making the judgment. In all cases, sound was off during facial affect coding, which was conducted in separate passes after gesture and protophone events had

already been coded. As with gaze direction, we eliminated from the analysis all "can't see" codes (coder unable to see the face during the gestural or vocal event). There were relatively few can't see codes for gestures (~6%), while protophones yielded more can't see judgments (~11%). The results showed a majority of both protophones and gestures yielded a neutral code for facial affect. There was a much greater tendency for facial affect to deviate from neutral (yielding a positive or negative judgment) for protophones (41%) than for gestures (16%). One possible interpretation of this tendency is that protophones tended to be used more emotionally expressively than gestures in the first year of life.

Data on Gaze Direction

The coders very rarely judged gestures as "can't see" with regard to gaze direction (<1%), while for protophones "can't see" was a much more common judgment (~10%). Social gestures coincided with person-directed gaze in 19.6% of instances, while non-social gestures coincided with person-directed gaze in 14.0% of instances (these values were calculated from grand totals across all ages and infants). It is important to consider these percentages in light of the definitions we used: non-social refers to gestures that are similar to vocal babbling in that they have no obvious social import in and of themselves (hand banging, foot shaking, …), while social gestures are ones that have naturally or conventionally-determined functions (reaching, offering, showing, nodding, waving…). Having a function that is determined does not, however, require that the action be intended as a communication by the child, only that observers could interpret the act (if they noticed it) *as if* it had been intended to serve a particular communicative function.

In subsequent research, we view it as important to determine gaze direction of the other participants in the recordings, especially the parents.

Data on Overlap of Gestures and Protophones

We evaluated overlap by reference to protophones that were adjacent in time to gestures, observing every case where a protophone overlapped in time with a gesture. Thus, we considered cases of temporally adjacent gestures and protophones, either 1) where a gesture was followed by a protophone (rather than another gesture), or 2) a protophone was followed by a gesture (rather than another protophone). There were 704 such adjacencies, 17% of which showed overlap between the protophone and the gesture. Presumably because protophones occurred so much more frequently than gestures, only 3% of protophones showed overlap with a gesture.