Use Authorization

In presenting this thesis in partial fulfillment of the requirements for an advanced degree at Idaho State University, I agree that the Library shall make it freely available for inspection. I further state that permission to download and/or print my thesis for scholarly purposes may be granted by the Dean of the Graduate School, Dean of my academic division, or by the University Librarian. It is understood that any copying or publication of this thesis for financial gain shall not be allowed without my written permission.

Signature _____

Date _____

Running head: TRANSCRIPTION IN VOCAL DEVELOPMENT

Phonetic Transcription in Vocal Development: When is Reliability Achieved?

Kayla Schroeder

A thesis

submitted in partial fulfillment

of the requirements for the degree of

Master of Science in the Department of Communication Sciences and Disorders

Idaho State University

August 2016

Committee Approval Page

To the Graduate Faculty:

The members of the committee appointed to examine the thesis of Kayla Schroeder find it satisfactory and recommend that it be accepted.

_____

Dr. Heather Ramsdell-Hudock,
Major Advisor

_____

Dr. Kristina Blaiser,
Committee Member

_____

Dr. David Mercaldo,
Graduate Faculty Representative

Idaho State
UNIVERSITY

Office for Research Integrity
921 South 8th Avenue, Stop 8046 • Pocatello, Idaho 83209-8046

September 1, 2015

Kayla Schroeder
Comm Sci Disorders/Deaf Educ
MS 8116

RE: regarding study number IRB-FY2016-45: Phonetic Transcription in Vocal Development: When is Reliability Achieved?

Dear Ms Schroeder:

I agree that this study qualifies as exempt from review under the following guideline: Category 4: Collection or study of existing data. This letter is your approval, please, keep this document in a safe place.

Notify the HSC of any adverse events. Serious, unexpected adverse events must be reported in writing within 10 business days.

You are granted permission to conduct your study effective immediately. The study is not subject to renewal.

Please note that any changes to the study as approved must be promptly reported and approved. Some changes may be approved by expedited review; others require full board review. Contact Tom Bailey (208-282-2179; fax 208-282-4723; email: humsubj@isu.edu) if you have any questions or require further information.

Sincerely,

Ralph Baergen, PhD, MPH, CIP
Human Subjects Chair

Phone: (208) 282-2179 • Fax: (208) 282-4723 • www.isu.edu/research

TABLE OF CONTENTS

List of Figures

List of Tables

Phonetic Transcription in Vocal Development: When is Reliability Achieved?
Thesis Abstract--Idaho State University (2016)

Phonetic transcription has been used to document infant vocalizations; however, research questions the reliability of this method. Accordingly, a need exists for a research base indicating a developmental age when transcription is reliable. This will allow time and resources to be preserved by utilizing other methods of documentation for younger children. The purpose of this study was to attempt to identify this age. Specifically, for a cohort of 7 infants, longitudinally gathered vocalizations from 7 to 18 months of age were transcribed by 3 coders who were intensively trained in phonetic transcription. Transcriptions were analyzed using an automated weighted reliability measure. It was hypothesized that transcription would be a reliable method to document vocalizations as infants approach 18 months of life. The results demonstrated increased reliability from 15 to 18 months of age when compared to reliability at younger ages. Clinical implications, study limitations, and future directions will be discussed.

Introduction

Researchers across the country are exploring caregiver report as it relates to infant vocal development. In doing so, more traditional approaches for tracking development, such as phonetic transcription, are being set aside. In order to maintain rigor in developing new caregiver report methods, we must justify reasoning for abandoning phonetic transcription at young ages. Additionally we must determine when phonetic transcription can be reinstated as an approach to document speech sounds in practice/research. Ultimately, both caregiver report and phonetic transcription are likely to play an important role in documenting speech sounds at some point in development, but we must conduct research to determine when.

Phonetic transcription has historically been used to document infant vocalizations; however, research calls into question the reliability and validity[1] of this method when used with prelinguistic sounds (Cucchiarini, 1996; Ramsdell, Oller, & Ethington, 2007; Ramsdell, Oller, Buder, Ethington, & Chorna, 2012; Stockman, Woods, & Tishman, 1981). Consequently, the ability to accurately identify young children with speech and

[1] Reliability and validity are often discussed in conjunction with each other. Inter-rater reliability is when similar results are produced by two coders for the same task. This is a good method to test phonetic transcription for infant vocalization because we are able to compare the similarity or difference of two sets of transcriptions for the same vocalization. With respect to the validity of transcription for infant vocalization, we must consider construct validity. Construct validity is a test of whether or not a procedure measures what it claims to measure. For phonetic transcription of infant speech, an utterance may contain a sound that does not have a phonetic symbol associated with it due to the immature nature of the vocalizations. As such, phonetic transcription may not be a tool that can accurately measure infant sounds, indicating a lack of validity. Accordingly, Ramsdell and colleagues (2007) discuss how transcription of infant sounds is hard to test for construct validity due to the inability to determine whether or not a transcription is accurate. Due to this inability to determine validity, inter-rater reliability will provide a better measure of whether phonetic transcription should be used at certain ages.

language delays utilizing transcription may also be hindered. A number of methods to increase the accuracy of phonetic transcription for older children and adults have been suggested (Knight, 2010; Louko & Edwards, 2001; Shriberg, Kwiatkowski, & Hoffmann, 1984); however, these methods are costly and time consuming, and the increased reliability is not likely to apply to infant vocalizations. Several alternative methods to transcription for infant vocalizations have also been introduced due to a perceived lack of accuracy (Ramsdell et al., 2012; Serkhane, Schwartz, Boë, Davis, & Matyear, 2007; Xu, Richards, & Gilkerson, 2014).

As evidenced by the number of alternative methods and published research suggesting a lack of transcription agreement[2] for infant vocalizations, there is a need for a research base indicating the developmental age when transcription can be used reliably. By identifying such an age, time and resources can be preserved, and alternate methods for documenting infant sounds at earlier ages can be developed. We propose the use of caregiver report of infant vocalizations as an accurate and cost effective method to use prior to the age when transcription is shown to be reliable (Ramsdell et al., 2012). Preliminary research has suggested an increase in transcription reliability at 20 and 21 months of infant age (Stockman et al., 1981). After finding low inter-judge and intra-judge reliability with criteria for identically matching transcribed segments, a stop feature

[2] For the sake of clarity, the terms "transcription reliability" and "transcription agreement" will be used interchangeably within this paper. However, as described by Cucchiarini (1996), agreement is technically a subtype of reliability. Additionally, the term "transcription accuracy" is defined as how well a transcription matches its target, or the "correct-ness" of a transcription. This is different than reliability or agreement in that transcription accuracy judges the precision of the transcription in relation to the target, rather than to the amount of similarity or difference between two transcribers (or one transcriber at different points in time).

analysis was applied, resulting in higher reliability (Stockman et al., 1981). Later research has expanded this analysis to a weighted reliability model based on sound features (Cucchiarini, 1996; Oller & Ramsdell, 2006). The stop feature analysis and weighted reliability models will be described in more detail below. Despite these reliability models, a specific infant age at which transcription reliability may be achieved within or across transcribers has not yet been identified.

Stages of Infant Vocal Development and Utterance Canonicity

In order to analyze how the transcription of infant vocalizations changes over time, it is helpful to have an understanding of infant vocal development. It has often been thought that children learn speech and language by imitation of adult speech. Research by Leonard, Fey, and Newhoff (1981) found two factors that affect children's imitation of words in relation to word learning: their linguistic level and their previous exposure to target words. At earlier linguistic levels the children's productions were tied to semantic knowledge, rather than the particular sounds of target words. When their knowledge of the target words increased, speech sound accuracy also increased. On the other hand, Howard and Messum (2011) argued that instead of infants learning speech by purely imitating their caregivers, speech is learned in an alternate manner. Infants play with articulation of sounds and learn different combinations; in turn caregivers imitate those infant utterances believed to contain well-formed structure or meaning. This interaction gives feedback to infants about which sounds are used in the ambient language, and eventually infants will discontinue the production of other sounds caregivers are not repeating back. While a specific infant age when this occurs was not listed, the researchers used a computational model of an infant that mimicked the vocal stages laid

out by Oller (2000): the phonation, primitive articulation, expansion, canonical, and integrative stages of vocal development. Both of the above studies are important in relation to transcription of infant vocalizations because the linguistic level, or stage of vocal development, is important to consider when transcribing infant speech, such that the more advanced linguistic level of the vocalization, the more reliable transcription will be.

Research on functional flexibility, the ability to produce one type of vocalization to represent different emotional meanings, has shown that infants as early as in the first year of life use this skill (Oller et al., 2013). This impacts future language learning because emotion is frequently used to both produce and perceive vocal messages. Additionally, these researchers noted that protophones including squeals, vowel-like vocalizations, and growls are not easily determined to be distinguishable consonants or vowels. Therefore, it is difficult to transcribe these sounds using the International Phonetic Alphabet (IPA). Rather than transcription, they suggest using behavioral categorizations (e.g., squeal, growl, etc.) before the infant enters the canonical babbling stage. Furthermore, Oller and Ramsdell (2006) noted that the IPA was formulated to document fully mature speech sounds (adult speech), with infant vocalizations not falling into this realm. Until the majority of an infant's utterances include more adult-like vocalizations, transcription will not accurately meet the need of documenting those sounds (Ramsdell et al., 2007).

According to Oller's stages of development (2000), the phonation stage occurs during the first 2 months of life and includes quasivowels and vegetative sounds (e.g., cries, burps, sighs, etc.). Quasivowels are inherently not well-formed, and are often

produced with the vocal tract at rest so they can be classified as fuzzy and imprecise, and therefore result in unreliable transcription. The primitive articulation stage occurs between 2 and 4 months of age and includes gooing, or primitive consonant-like sounds accompanied by quasivowels. Again, these sounds are not well-formed and will lead to imprecise transcriptions. The next stage is the expansion stage, which occurs between 4 and 7 months of age and is characterized by the use of fully resonant nuclei (i.e., well-formed vowels), squealing, growling, whispering, yelling, raspberries, and marginal babbling. At this point in vocal development, we may begin to see a minimal increase in transcription reliability given more articulate vowel productions, however consonant sounds are still immature, and vowel/consonant combinations occur within prolonged timeframes resulting in continued difficulty identifying accurate phonemes in productions.

As Oller (2000) indicated, infants then proceed to the canonical stage of vocal development where they begin to truly babble. This stage typically begins with reduplicated babbling, or using the same syllable two or more times in a row (e.g., “baba,” “mamama,” etc.), and progresses to variegated babbling, or combining different syllable shapes in different orders (e.g., “mibadee,” “botika,” etc.). The unique characteristics of canonical syllable shapes that are likely to lead to additional increases in transcription reliability are the clearly articulated vowel and consonant segments, with timely transitions between the two. Given these characteristics, canonical syllables are the first well-formed productions, matching syllables in the ambient language. Canonical babbling can begin as young as 7 months and continues into the next stage, the integrative stage. The integrative stage is defined as the combination of canonical

babbling, protowords, and true words. Early words are often formulated using reduplicated or variegated babbling (each made up of canonical syllables), for example: “baba” for “bottle” and “baykee” for “blanket.” Additionally, children in this stage commonly babble without meaning in combination with the words they utilize that contain meaning. This stage begins around 12 months with the emergence of protowords and the first true words, and continues to about 18 months of age. In the current study, we are most interested in the integrative stage because it is thought that in this stage infant vocalizations will be more well-formed as the majority of vocalizations contain canonical syllables, protowords, and/or true words.

Additionally, research done by Ramsdell and colleagues (2007) found higher reliability for vocalizations that contain canonical syllables. As discussed above, it is expected that more adult-like vocalizations, such as canonical babbling and early words, will be present in the first half of the second year of life (Oller, 2000). Accordingly, we expect transcription of infant vocalizations to be more reliable as the infant enters that age range, creating a basis for the further exploration of infant age of transcription reliability in this study.

Measuring Transcription Reliability

Several methods have been published for documentation of transcription reliability. In a study conducted by Stockman and colleagues (1981) on the reliability of transcription for infant vocalizations, they applied a stop-feature analysis after finding low inter and intra-transcriber reliability for identically matched segments. The original identical matching criteria began by lining up comparable segments of differing transcriptions and counting segments that were identically transcribed as being in

agreement. During the process of lining up comparable segments the researchers noticed that some segments only varied by voicing. For instance, the transcribed segments /p/ and /b/ have the same place and manner production features; however, they differ in that /p/ is voiceless and /b/ is voiced. If transcribed for the same segment in the original analysis, these two sounds would have been marked as not being in agreement with a reliability score of zero. Because of the high frequency of stops (/p, b, t, d, k, g/) in comparison to other voiced/voiceless cognates in infant vocalizations, the researchers introduced the stop-feature analysis. In this analysis, matched segments containing oral stops were given a reliability score of one regardless of voicing classification.

The stop-feature analysis was later expanded, in other research, to an agreement index, allowing for alignment of sounds based on the similarity of articulation, or multiple phonetic features beyond just voicing (Cucchiarini, 1996). The expanded agreement index used 10 features for consonants, 3 features for vowels, and allowed for the inclusion diacritic markers. The features used to compare the degree of similarity for consonants were: place, voice, nasality, stop, glide, lateral, fricative, trill, height, and distribution. The features used to compare the degree of similarity for vowels were: front/back, tongue height, and lip rounding. Diacritic markers were important for the analysis because they could affect the similarity or difference of two sounds. This is shown by the following example: if [f] and [s] were transcribed by two different transcribers for the same segment, the addition of the dentalized diacritic to the [s] would make the two transcriptions more similar than if dentalization was not included. The dentalized diacritic would indicate that the voiceless alveolar fricative /s/ was produced in a more anterior place of articulation. Given that the voiceless labiodental fricative /f/ is

naturally produced more anteriorly than /s/, an advanced place of articulation for /s/ would make the two transcriptions more similar.

Oller and Ramsdell (2006) argued for a weighted model that could more precisely document reliability of transcribed samples from speakers of different languages and across different ages. This weighting was achieved by attributing three types of agreement to a set of transcriptions for the same vocalization: global structural agreement, featural agreement, and overall transcription agreement. Global structural agreement is when a segment was given a value of either one or zero: one if the two transcriptions both had an assigned transcription for a certain segment, and zero if one of the transcriptions did not have an assigned transcription for a segment resulting in an orphan segment. Featural agreement is how similar or different two matched segments were in relation to phonetic features: with identically transcribed segments obtaining a score of one, those with no shared phonetic features receiving a score of zero, and those with some shared features receiving a scaled score between one and zero based on the degree of similarity. The overall transcription agreement was calculated by multiplying the global structural agreement with the featural agreement.

Using this weighted model, Oller and Ramsdell (2006) analyzed samples from an infant, a toddler, and three adults: Korean-speaking, Ukrainian-speaking, and American English-speaking. These samples were transcribed, aligned by researchers, and assessed for inter-transcriber reliability with a weighted reliability computer program. Results of the study indicated increased reliability of transcriptions for all samples with the weighted measure when compared to an unweighted measure, or strict match criteria. Utilizing the weighted measure, the average transcription agreement was approximately

0.6 for the infant sample, 0.75 for the toddler sample, 0.8 for the Korean adult sample, 0.82 for the Ukrainian adult sample, and 0.9 for the English adult sample. The unweighted measure produced reliability values of approximately 0.2, 0.4, 0.45, 0.55, and 0.65 respectively for the above samples. The weighted reliability measure was shown to have higher reliability because with the traditional unweighted measure (or strict match criteria) there is often a floor effect which limits the amount transcriptions can be compared. While floor effects can impact any unweighted reliability analysis, they were especially observed in relation to transcription of the infant sample. This was demonstrated by a high number of zero agreement values found in the infant sample; using the unweighted measure, 103 out of 210 segments compared for reliability were given zero values, or a lack of complete agreement. It follows that while using an unweighted measure, these segments were defined as having no agreement whatsoever; however, the majority of the comparable segments matched as either a consonant or vowel and/or had similar features (e.g., both fricatives). This floor effect prevented any similarities in transcription to be compared by instead giving a zero value for segments that might only differ slightly.

Further research has expanded the weighted reliability model to be weighted for the accuracy of a production when compared to a correct target word (Preston, Ramsdell, Oller, Edwards, & Tobin, 2011). This research has implications in regards to providing better reliability across different speakers and levels of well-formedness. In the study, Preston and colleagues (2011) found that the weighted measure could accurately identify speech disorders, which relate to infant vocalizations because those with speech disorders often have less well-formed productions.

Methods to Increase Transcription Accuracy

Methods to increase the accuracy of phonetic transcription have been identified by various researchers. Louko and Edwards (2001) compiled several methods to aid in enhancing transcription accuracy for clinical practice, especially in relation to unintelligible speech, or otherwise difficult to transcribe samples. Because prelinguistic vocalizations often fall into the unintelligible realm and rarely represent actual words, these methods can be applied to infant vocalizations. However, as discussed previously, it is difficult to determine whether the transcription of an infant vocalization is accurate because we cannot determine what the target production is. Therefore, when applying these methods to infant speech one must hope that there is an increase in accuracy which will also cause an increase in reliability. One such method is to have multiple coders discuss the transcription of difficult words and vocalizations in order to obtain better perspective and come to an agreement on the transcription (Louko & Edwards, 2001). Transcription by consensus can also be conducted, with two or more transcribers independently transcribing a segment, and then comparing the transcriptions while listening to the production again in order to determine the true transcription (Shriberg et al., 1984). Another way to increase transcription accuracy is to recognize common error patterns; either those frequently made by the client (such as a client who frequently adds the phoneme /b/ after /m/), or those common in typical developmental (such as the substitution of /w/ for /r/ before the age of 6; Louko & Edwards, 2001). By recognizing such patterns, the transcriber may more easily determine the correct transcription of the production if it is necessary to transcribe on line (i.e., at the time the production is happening).

An important method for increasing the accuracy of transcription is to either audio or videotape the sample being transcribed, rather than transcribing on line (Louko & Edwards, 2001). When attempting to transcribe on line, sounds or words are often missed due to the speed of the interaction. Recording the sample provides an opportunity to listen to the sample multiple times, therefore increasing the potential for transcription accuracy (Knight, 2010; Louko & Edwards, 2001). Knight (2010) conducted a study with undergraduate students in a phonetics class acting as transcribers. The students listened to nonsense words produced by two different speakers. Results indicated that overall accuracy was higher after listening to a nonsense word 10 times, over listening to the same nonsense word six times. Louko and Edwards (2001) detailed a methodology for repeated listening: first listen to the whole word or vocalization, then listen in detail to each syllable or sound segment. Hold a "mental template" (p. 7) of the sound or sounds heard while listening to the segment again to see if the audio matches the imagined template. Once every segment has been transcribed, listen to the whole word or vocalization to see that no part of the word was ignored and all of the corresponding segments go together in the whole vocalization.

Other methods for increasing the accuracy of transcriptions pertain to how a transcription sample is analyzed, or the type of transcription used. Vihman (1986) obtained samples of children at 1 and 3 years of age, and analyzed the transcriptions for patterns of sound classes (e.g., fricatives, stops, etc.) and syllable types (e.g., syllables containing only a vowel vs. syllables starting with a consonant and containing a vowel, and other combinations of consonants and vowels), rather than just using the raw transcription data. The purpose of this analysis was to attempt to predict phonological and

language ability at age 3 by analyzing babbling output at age 1. Results of this study indicated that while specific sound classes and types of articulation did not predict ability at age 3, overall consonant use at age 1 predicted how advanced phonological ability was at age 3. Shriberg and Lof (1991) compared broad and narrow transcriptions for inter and intra-transcriber reliability using 51 subjects from a diverse group which included: children who were typically developing, children who were developmentally delayed, children with disordered speech, and five adults with intellectual disability. Broad transcription is a more general type of transcription where only the phoneme produced is transcribed. Narrow transcription takes more detail into account and uses diacritic markers to note slight differences in production. The researchers found that while inter-judge and intra-judge reliability were equivalent in terms of agreement percentages, broad transcription resulted in significantly higher reliability than narrow transcription (difference = 19%). This indicated that broad transcription should be used in most cases due to its relatively high reliability when compared to narrow transcription. The complication here, however, is that narrow transcription provides more information about a production than broad transcription.

Alternate Methods to Transcription

Due to a perceived lack of accuracy when transcribing, especially in regards to transcription of infant vocalizations, researchers have begun testing and using alternative methods to document infant vocal development. Transcription of infant vocalizations often takes a great deal of time because of the ambiguity of infant vocalizations. Researchers must take this time themselves, or train laboratory staff on transcription of infant vocalizations. If transcription of infant vocalizations is shown to not be reliable

before a certain age, other methods can be utilized to document vocal development before that age, therefore saving time and money. One alternate method that has been introduced is the identification of prelinguistic phonological categories using caregiver report of vocalizations (Ramsdell et al., 2012). Researchers asked caregivers to report what sounds their infants were making, and compared caregiver report to a naturalistic listener in the lab, and to transcription of the infant sounds based on monthly lab recordings. They found significantly higher repertoire sizes when the vocalizations were transcribed as compared to when caregivers and naturalistic listeners reported on the sounds. However, higher repertoire sizes do not necessarily equate to better accuracy; on the contrary, these results suggest that caregivers notice salient patterns in vocalizations, or those production abilities that are useful in guiding word learning and language development. Transcribers, on the other hand, attempt to give significance to every utterance, making transcription a less accurate measure of the infant's functional abilities. Caregiver report is natural, does not require special training, and is cost-effective. We propose the use of this method before the age when transcription is shown to be reliable because it can easily be implemented in both clinical practice and research, and is supported by research.

Furthermore, there are several other methods that have been introduced to document infant vocal development, and potentially circumvent transcription. At 4 and 7 months of infant age, Serkhane and colleagues (2007) used an articulatory acoustic model to compare formant values (F1 and F2) to the place of articulation in order to describe infant vocalizations as the vocal tract grows. This was achieved by using a computerized model based on the size and range of motion of an infant's vocal tract. They matched the model to a corpus of vocalizations from 24 infants at 4 months of age and 3 infants at 7

months of age. By doing this, the researchers were able to see the jaw movement, tongue movement, larynx height, and lip shape required to produce certain vocalizations at different infant ages. Another alternate method to transcription that has been explored in research is an automated computational measure to segment sounds and analyze them based on 39 English phonemes used in adult speech (Xu et al., 2014). This method enables identification of group differences and possible identification of delays/disorders as evidenced by significant group differences between 106 children who were typically developing, 71 children who had autism, and 49 children who had a learning disability (not related to autism). Although alternative, the level of training required to implement the above two methods would be cumbersome, and the likelihood of clinicians finding them useful for identifying infants with delayed/disordered speech and/or language is slim.

Preliminary Evidence of Infant Age of Transcription Reliability

Preliminary research on infant age when transcription becomes reliable has suggested an increase in reliability at 20 and 21 months of infant age (Stockman et al., 1981). Stockman and colleagues transcribed the vocalizations of four infants from recordings periodically obtained between 7 and 21 months of age. The results indicated relatively low (less than 60%) inter and intra-judge reliability across infant ages for matched transcribed segments. Reliability only dramatically increased at 20 and 21 months of age for transcriptions of one infant, presumably caused by an increased use of true words. After finding such low reliability for identically matched segments, their stop feature analysis (described above) was applied, which resulted in higher reliability.

Despite previously published findings, a specific infant age when transcription becomes reliable has not yet been determined.

Goals and Rationale

The *long-term goal* of this research is to identify an age when transcription can be reliably used to document infant vocalizations, in order to enable other methods, such as caregiver report, to be utilized at younger ages. The *objective* of this study is to provide preliminary evidence of a specific age when transcription of infant vocalizations becomes reliable using a weighted reliability model. The *central hypothesis* is: given infants who are typically developing, it is predicted that transcription reliability will increase with infant age. This hypothesis is based on knowledge of infant vocal development and research showing canonical syllable transcription reliability (Ramsdell et al., 2007). Additional support for this hypothesis comes from previous research, which indicated an increased ease of transcription when recognizable words were present in vocalizations (Stockman et al., 1981). The *rationale* for the proposed research is that, once a strong data set demonstrating increased reliability of transcription starting at a certain age is established, methods such as caregiver report can be developed for use before that age.

The central hypothesis of this project was tested by pursuing the following *aim:* from 7 to 18 months of infant age, across typically developing infants, we identified inter-transcriber reliability patterns using a weighted reliability measure. Based on prior documentation that canonical utterances produce higher transcription reliability (Ramsdell et al., 2007), the *working hypothesis* for this aim was that higher inter-judge transcription agreement would occur between 15 to 18 months of age, when the majority of utterances are canonical and/or contain early word forms.

Methods

Participants

Vocalizations for this study were obtained from seven infants video/audio recorded monthly in a study conducted by Dr. Heather Ramsdell-Hudock at East Carolina University (ECU). All infants were 6 and 18 months of age at the beginning and termination of the study, respectively. For the purposes of this project, we explored data from 7 through 18 months of infant age. Following previous approval from the University Medical Center Institution Review Board at ECU, caregivers voluntarily gave informed consent for participation in the study. Further, exemption was obtained from the Human Subjects Committee at Idaho State University (ISU), as the purpose of the proposed study is covered in the original consent.

All families were of middle socioeconomic status (as determine through parent self-report on participant history interview). There were no infant participants born to single parent homes, and both mothers and fathers participated in the original study. Four of the infants were first born, one had one older sibling, one had two older siblings, and one had three older siblings. Siblings ranged in age from 2 years to 5 years at the time of infant participants' births.

Three of the seven infant participants were male, and four were female. One female infant was African American, and one male infant was Palestinian. The male infant who was Palestinian was from a home where English and Arabic were spoken. All infants had normal hearing; they all passed an automated auditory brainstem response newborn screening (ALGO 3 or ALGO 5 Newborn Hearing Screener System) to click stimuli presented at 35 dB nHL. In addition, full hearing evaluations including

tympanometry, transient evoked otoacoustic emissions, and visual reinforcement audiometry were conducted at 6 and 18 months of age, with follow-up testing as needed for instances where results were abnormal (i.e., middle ear dysfunction) or testing was incomplete. One of the infants received bilateral myringotomy and pressure equalization tubes during their enrollment in the study. Anecdotally, regardless of language background or hearing status, all infants demonstrated typical speech and language development during the recording period.

Procedure

Data from the recordings was prepared by trained laboratory staff, who located infant utterances based on a breath group criterion (excluding both vegetative noises and utterances with substantial overlay from another noise source; Oller & Lynch, 1992). Located utterances were extracted from the original recording file and randomly selected for transcription. Three transcribers, intensively trained in using the IPA, followed strict protocol in transcribing: working independently, not viewing acoustic displays, listening to each utterance no more than six times, and including exotic sounds in transcriptions that may not be part of the phonemic repertoire of General American English. Phonemes used in transcriptions of these vocalizations included those listed in Tables 1 and 2. The rationale behind limiting the number of times a transcriber may listen to an utterance is that due to the imprecise nature of infant vocalizations, a person could possibly hear a different sound after listening to it multiple times. Given that coders often report low confidence in transcription of infant vocalizations, they could potentially listen to an utterance for an extensive period of time. By limiting the amount of times vocalizations

can be listened to, the transcriber is forced into a decision between two or more sounds they are unsure about, which helps save time in the transcription task.

All transcriptions were systematically aligned in accordance with principles set forth in previous research (Oller & Ramsdell, 2006; Ramsdell et al., 2007), including 4 alignment principles. The first principle is the *strict-order principle,* which requires that all segments in a

Table 1. Pulmonic consonants that were used in this study (bolded symbols represent sounds that occur in English)

	Bilabial	Labio-dental	Dental	Alveolar	Post-alveolar	Retro-flex	Palatal	Velar	Uvular	Pharyn-geal	Glottal
Stop plosive	**p b**				**t d**			**k g**	q ɢ		**ʔ**
Nasal	**m**			**n**		ɳ		**ŋ**	ɴ		
Trill	B			r					ʀ		
Tap or Flap				**ɾ**							
Fricative	ɸ β	**f v**	**θ ð**	**s z**	**ʃ ʒ**		ç ʝ	x ɣ	χ ʁ	ħ ʕ	**h** ɦ
Lateral Fricative				ɬ ɮ							
Approximant		ʋ		ɹ		**ɻ**	**j**				
Lateral Approximant				**l**			ʎ	ʟ			

Symbols to the right in a cell are voiced, and to the left are voiceless.

Table 2. Vowels used in this study (bolded symbols represent sounds that occur in English)

	Front	Central	Back
Close	**i** y	ɨ ʉ	ɯ **u**
	ɪ ʏ		**ʊ**
Close-mid	**e** ø	ɘ ɵ	ɤ **o**
		ə	
Open-mid	**ɛ** œ	ɜ ɞ	**ʌ** **ɔ**
	æ	ɐ	
Open	**a** ɶ		**ɑ** ɒ

Symbols to the right in a pair are rounded.

transcription remain in their original order. The *matched segment principle* requires vowel-like and consonant-like segments that are in the same order to be matched together

in aligned transcriptions. Next, when there are different numbers of vowel-like or consonant-like segments, or those segments are not ordered in the same way, the *minimum discrepancy principle* calls for the alignment of segments in such way as to create the most phonetically similar segment matches without reordering any segments as discussed in the *strict-order principle*. Finally, the *nucleus alignment first principle* requires vowel-like segments to be aligned first because of the perception of vowels as the center of a syllable.

Once aligned, weighted reliability between transcriptions was calculated by a program written in LIPP™ (Logical International Phonetics Programs) analysis language (LAL; Oller & Delgado, 1999). Weighting was achieved by comparing the segments from two aligned transcriptions. Each aligned segment of a vocalization, as transcribed by two transcribers, was weighted on a scale from 0-1, based on how similar or different the phonetic features of the transcriptions were. Three types of agreement are used in this weighted measure: global structural agreement, featural agreement, and overall transcription agreement (Oller & Ramsdell, 2006). Global structural agreement is when each set of aligned segments is given a value of either one or zero: a value of one if each transcription contains a segment in a similar position after alignment, and a value of zero if one of the transcriptions does not contain a segment, which would result in an orphaned segment. Featural agreement is the similarity or difference of phonetic features on a scale of zero to one of two matched segments. Overall transcription agreement is calculated by multiplying the global structural agreement with the featural agreement.

Consider the following example:

Coder A	[p	ĩ	n	]
Coder B	[b	i	d	i]

In this example, the first three segments would have a global structural agreement of 1 and the last segment would have a global structural agreement of 0 because the Coder A did not transcribe a segment and Coder B did. The mean of all global structural agreements for each segment is the global structural agreement for the entire vocalization. In this example the global structural agreement would be 0.75 for the whole vocalization. For featural agreement, the difference between the first segments of the two transcriptions is a voicing difference, which gives the two segments a featural agreement of 0.67. In the second segment, the same symbol is transcribed in both with the nasalization diacritic on the top transcription. This small difference produces a featural agreement of 0.9. In the third segment both transcriptions have the same place of articulation, but a different manner of articulation, again resulting in a featural agreement of 0.67. The mean featural agreement of the three slots that have segments transcribed by both coders is 0.74. To obtain the overall transcription agreement we multiple the mean featural agreement and the mean global agreement to get 0.56.

Analysis

A repeated-measures analysis of variance (ANOVA) was used to evaluate the variables of interest. The dependent variable of interest was phonetic transcription reliability, as calculated through the weighted reliability measure in LIPP™. The independent variable of interest was infant age (monthly from 7 through 18 months of age).

Results

A repeated measures ANOVA with sphericity assumed showed that weighted transcription reliability differed statistically significantly across infant ages [F (11, 66) = 3.432, $p < 0.001$), as displayed in Figure 1. Post hoc tests using Tukey's LSD revealed that transcription reliability was statistically significantly lower at 7 months (M = 0.517, SD = 0.061) than at 15 (M = 0.657, SD = 0.078, p = 0.005), 16 (M = 0.606, SD = 0.037, p = 0.009), 17 (M = 0.634, SD = 0.051, p = 0.016), and 18 (M = 0.623, SD = 0.025, p = 0.009) months of infant age. Additionally, transcription reliability was statistically significantly lower at 8 months (M = 0.546, SD = 0.056) than at 17 (p = 0.006) and 18 (p = 0.014) months of infant age; at 9 months (M = 0.556, SD = 0.066) than at 18 (p = 0.025) months of infant age; at 10 months (M = 0.571, SD = 0.061) than at 18 (p = 0.016) months; and at 11 months (M = 0.536, SD = 0.080) than at 18 months (p = 0.027) of infant age. Therefore, we can conclude that infant age, and production of more canonical and linguistic vocalizations, elicits a statistically significant increase in weighted transcription reliability, but only after 15 to 18 months of infant age will significant increases be observed.

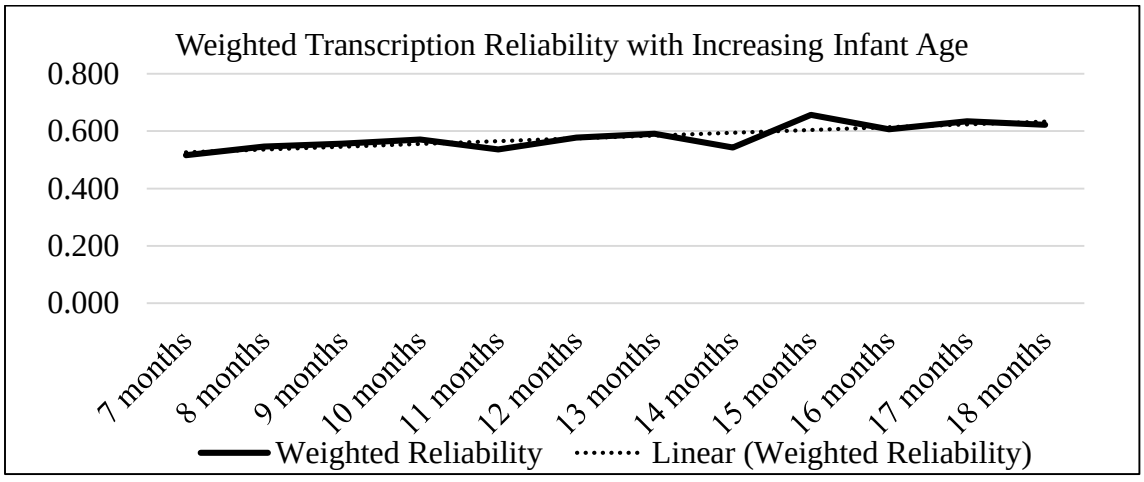

Figure 1. Weighted transcription reliability from 7 through 18 months of infant age.

Further, given that there were three coders (with each infant transcribed twice - Coder A transcribed all of the infants, Coder B transcribed four of the seven infants, and Coder C transcribed the additional three infants - such that Coder A's transcriptions were compared with Coder B's and C's), an independent-samples *t*-test was conducted to compare the weighted transcription reliability between Coder A and B and between Coder A and C. The comparison of the weighted transcription reliability with increasing infant age across these Coder pairings can be viewed in Figure 2. There was not a statistically significant difference in the weighted transcription reliability between Coder A and B ($M = 0.598$, $SD = 0.002$) and between Coder A and C ($M = 0.569$, $SD = 0.002$), $t(10) = 2.092$, $p = 0.063$. These results suggest that there were no substantive differences between the coders who transcribed the infant vocalizations for this study.

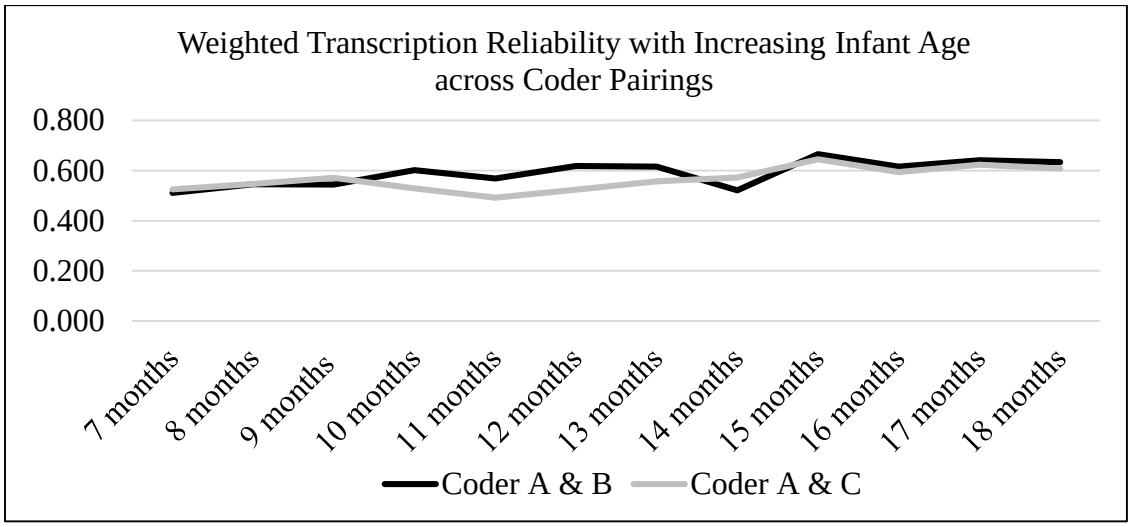

Figure 2. Weighted transcription reliability across coders (Coder A compared with Coders B and C) from 7 through 18 months of infant age.

Discussion

The purpose of this study was to explore phonetic transcription reliability for tracking infant vocal development. In doing so, we hoped to determine an age at which transcription can be used reliably in clinical and research settings. Specifically, through this study, we analyzed inter-coder reliability for transcription of 7 infants' vocalizations at each month from 7 to 18 months. Our hypothesis was shown to be correct, that transcription reliability increases with infant age and is higher between 15 to 18 months of age. Using a weighted reliability measure, our results indicated that there is a significant difference in transcription reliability across infant ages. Specifically, reliability was shown to be statistically significantly higher at 15, 16, 17, and 18 months of age than at younger ages.

Mean inter-transcriber reliability values ranged from 0.517 to 0.657, which indicated increased reliability at later ages. Intuitively, a higher reliability is better than a lower reliability, but this brings up the question of how high of a reliability value is acceptable to determine an age with which transcription can be used. Lance, Butts, and Michel (2006) discuss a common citation that a reliability value of 0.7 is acceptable. However, the reference from Nunnally (as cited in Lance et al., 2006) is much more complex, and only says to use a reliability value of 0.7 for the early stages of research. Additionally, Nunnally (as cited in Lance et al., 2006) indicates that for applied research, a reliability of 0.8 is the lowest acceptable value, but 0.9 or 0.95 reliability is much more desirable. With the highest reliability value in the current study being 0.657, we do not even reach the commonly accepted reliability value of 0.7. This indicates that while this study has shown increasing reliability for phonetic transcription between 7 through 18

months of age, the reliability values we obtained do not suggest that transcription is an acceptable method for documenting vocal development at even the latest age explored. In future studies, we suggest a reliability value of at least 0.8 to indicate the ability to accurately use reliability at that age, due to the highly applied nature of transcription.

Our results were congruent with the results found in previous research (Oller & Ramsdell, 2006; Stockman et al., 1981). While Stockman and colleagues (1981) had results of less than 0.6 reliability for all ages in their study using an unweighted reliability measure, our results indicated some reliability above 0.6, but overall still below 0.7, using a weighted reliability measure. This indicates that while a weighted reliability measure produces higher reliability, having accounted for similar features across transcriptions, transcription of infant vocalizations still results in relatively low reliability whether using a weighted or unweighted measure. This was also demonstrated in Oller and Ramsdell's study (2006), where they found 0.6 weighted reliability for transcription of an infant in the canonical stage of development. Again, this suggests that transcription is not a reliable method for documenting infant vocalizations. Stockman and colleagues (1981) transcribed infants up to 21 months of age, whereas the current study only went up to 18 months. They found increased reliability with the unweighted measure for one infant at 20 and 21 months of age and overall higher reliability across ages using the stop feature analysis (a form of weighting) (Stockman et al., 1981). Additionally, Oller and Ramsdell (2006) found a reliability value of 0.75 for a sample of a 24 month old in their study. Given these increases, a future direction for research might be to use the weighted reliability method described in the current study for transcription of infants up to 25 months of age and older to determine if a transcription reliability of 0.8 can be achieved.

The increase in reliability seen in this study at later ages might be because of increased use of canonical syllables, protowords, and/or true words. Canonical syllables have been shown by Ramsdell and colleagues (2007) to produce higher reliability than non-canonical syllables. As infant age increases, there will hypothetically be more canonical syllables in addition to early words and true words. Anecdotally, it is expected that as a child begins to use more real words than babbling, transcription will be able to be used reliably. However, future research needs to address this with an expanded age range of children. Prior to this time, new methods of tracking vocal development need to be used. We propose the use of caregiver report of infant vocalizations as a reliable, time efficient, and cost-effective method to use prior to when transcription becomes reliable (Ramsdell et al., 2012).

Clinical Implications

As discussed above, inter-transcriber reliability of infant vocalizations was higher at 15 to 18 months of age, but still below the commonly accepted 0.7 reliability, and well below the more desired 0.8 reliability. This indicates that transcription is not an accurate measure of infant vocal development for clinical use at these ages. For instance, if a clinician was trying to determine what sounds an infant was making to track development it would be impractical to record and phonetically transcribe a certain amount of the infant's utterances when the reliability of that transcription would be relatively low. Additionally, it is not guaranteed that a sampling of the infant's utterances would include all of the sounds in the infant's repertoire.

By implementing the use of caregiver report of vocal development at these younger ages, the same clinician mentioned above can easily ask the child's caregiver

what sounds the infant is making. As shown in previous research, the caregiver is likely to provide a report of the infant's true repertoire, by noting sounds and words that are salient and predictive of later language use (Ramsdell et al., 2012). Additionally, this method is quicker and more cost effective than transcription (Ramsdell et al., 2012).

Study Limitations

A potential limitation to the study is that for some ages, a portion of the 30 randomly selected utterances to be transcribed from a few of the infants was reduced. This occurred when at least one of the transcribers classified an utterance as untranscribable (i.e., wrote "cry"), or it appeared that the extracted utterance was not an infant vocalization (i.e., background noise, a sound from a toy, etc.). Additionally we did not have data from all of the infants at all of the ages. If we had compared transcription of 30 utterances for each of the 7 infants at each of the 12 months we could have compared 2520 utterances. Due to these two reasons for reduction in utterances, we only compared 1967 utterances for transcription reliability, which could have altered the results of the study.

Another limitation to the study was that while all transcribers were instructed to include sounds that are non-native to English (i.e., a uvular fricative) as appropriate, there is the possibility for bias towards sounds that occur in English because all transcribers in this study use English as their primary language. Typically people are more likely to hear sounds that occur in their own language than those that occur in a non-native language. However, each coder had some language training in languages other than English (Spanish or Russian), which might have helped to reduce that bias toward English phonemes. In addition, each coder was accustomed to listening to the exotic quality of

infant vocalizations, having spent hours per week for multiple years working in the area of infant vocal development.

Another limitation to the current study is the relatively low sample size of seven infants, studied only through 18 months of age. A larger sample size of 16 infants could have been obtained from archived date, but the time needed to transcribe that number of utterances was unrealistic for the scope of a thesis project. The archived data only contained recordings through 18 months of age, so a new corpus of data would need to be obtained for further analysis at later ages. Additionally, we only analyzed inter-transcriber reliability (two different transcribers) and did not address intra-transcriber reliability (one transcriber at two different points in time). This was something that could not be achieved due to time constraints, but could be a direction for future research.

Future Directions

Based on the results and clinical implications, it is suggested that future research be conducted utilizing a larger sample size with infants through 25 months of age. This would allow for potential observation of when inter-transcriber reliability reaches higher than 0.8, which would determine when transcription could be used past that age. By continuing to explore phonetic transcription reliability values as infants age through 25 months, it is likely that clinicians and researchers will be able to save time and resources from being wasted on the use of an unreliable measure at early infant ages.

Conclusion

At this time we are unable to determine at what age phonetic transcription can be used to reliably document speech sound productions. Supporting past research, we have shown that phonetic transcription reliability increases with infant age from 7 through 18

months. However, reliability values obtained of 0.657 are not acceptable to determine an age when phonetic transcription can be used to document development. Expanding the infant age range through 25 months may enable researchers to indicate an age at which phonetic transcription can be used more reliably with this population. More research should be conducted on phonetic transcription reliability for infant vocalizations. A larger sample size of infants, as well as older ages should be examined for increased reliability. This information may be useful in helping clinicians and researchers to preserve time and resources by utilizing other methods, such as caregiver report, to document vocalizations for younger children.

References

Cucchiarini, C. (1996). Assessing transcription agreement: methodological aspects. *Clinical Linguistics & Phonetics, 10*(2), 131-155.

Howard, I. S., & Messum, P. (2011). Modeling the development of pronunciation in infant speech acquisition. *Motor Control, 15*(1), 85-117. Retrieved from http://eds.b.ebscohost.com/eds/pdfviewer/pdfviewer?sid=433a6bfc-c499-4356-8a6e-f8856b395b54%40sessionmgr112&vid=2&hid=127

Knight, R. (2010). Transcribing nonsense words: The effect of numbers of voices and repetitions. *Clinical Linguistics & Phonetics, 24*(6), 473-484. doi: 10.3109/02699200903491267

Lance, C. E., Butts, M. M., Michels, L.C. (2006). The sources of four commonly reported cutoff criteria: What did they really say? *Organizational Research Methods, 9*(2), 202–220. doi: 10.1177/1094428105284919

Leonard, L. B., Fey, M. E., & Newhoff, M. (1981). Phonological considerations in children's early imitative and spontaneous speech. *Journal of Psycholinguistic Research*, *10*(2), 123-133.

Louko, L. J., & Edwards, M. L. (2001). Issues in collecting and transcribing speech samples. *Topics in Language Disorders*, *21*(4), 1-11.

Oller, D. K. (2000). *The emergence of the speech capacity*. Mahwah, N.J.: Lawrence Erlbaum Associates.

Oller, D. K., Buder, E. H., Ramsdell, H. L., Warlaumont, A. S., Chorna, L., & Bakeman, R. (2013). Functional flexibility of infant vocalization and the emergence of

language. *Proceedings of the National Academy of Sciences*, *110*(16), 6318-6323. doi: 10.1073/pnas.1300337110

Oller, D. K., & Delgado, R. E. (1999). *Logical International Phonetics Program* (Version Windows). Miami: Intelligent Hearing Systems Corp.

Oller, D.K., & Lynch, M.P. (1992). Infant vocalizations and innovations in infraphonology: Toward a broader theory of development and disorders. In C. Ferguson, L. Menn, & C. Stoel-Gammon (Eds.), *Phonological development: Models, research, implications* (pp. 509-536). Parkton, M.D.: York Press.

Oller, D. K., & Ramsdell, H. L. (2006). A weighted reliability measure for phonetic transcription. *Journal of Speech, Language, and Hearing Research, 49*(6), 1391-1411.

Preston, J.L., Ramsdell, H.L., Oller, D.K., Edwards, M.L., & Tobin, S.J. (2011). Developing a weighted measure of speech sound accuracy. *Journal of Speech, Language, and Hearing Research, 54*, 1-18. doi: 10.1044/1092-4388(2010/10-0030)

Ramsdell, H. L., Oller, D. K., Buder, E. H., Ethington, C. A., & Chorna, L. (2012). Identification of prelinguistic phonological categories. *Journal of Speech, Language, and Hearing Research, 55*(6), 1626-1639. doi: 10.1044/1092-4388(2012/11-0250)

Ramsdell, H. L., Oller, D. K., & Ethington, C. A. (2007). Predicting phonetic transcription agreement: Insights from research in infant vocalizations. *Clinical Linguistics & Phonetics, 21*(10), 793-831.

Serkhane, J. E., Schwartz, J. L., Boë, L. J., Davis, B. L., & Matyear, C. L. (2007). Infants' vocalizations analyzed with an articulatory model: A preliminary report. *Journal of Phonetics, 35*(3), 321-340.

Shriberg, L. D., Kwiatkowski, J., & Hoffmann, K. (1984). A procedure for phonetic transcription by consensus. *Journal of Speech, Language, and Hearing Research, 27*(3), 456-465.

Shriberg, L. D., & Lof, G. L. (1991). Reliability studies in broad and narrow phonetic transcription. *Clinical Linguistics & Phonetics*, *5*(3), 225-279.

Stockman, I. J., Woods, D. R., & Tishman, A. (1981). Listener agreement on phonetic segments in early infant vocalizations. *Journal of Psycholinguistic Research, 10*(6), 593-617.

Vihman, M. M. (1986). Individual differences in babbling and early speech: Predicting to age three. *Precursors of early speech*, *44*, 95-109.

Xu, D., Richards, J. A., & Gilkerson, J. (2014). Automated analysis of child phonetic production using naturalistic recordings. *Journal of Speech, Language, and Hearing Research*, *57*(5), 1638-1650. doi: 10.1044/2014_JSLHR-S-13-0037