Monosyllabic Mandarin Tone Productions by Three-year-olds Growing up in Taiwan and the U.S.: Inter-judge Reliability and Perceptual Results

Puisan Wong

The Ohio State University

Abstract

Purpose

This study compared monosyllabic Mandarin lexical tones produced by three-year-old Mandarin-speaking children growing up in Taiwan and the U.S.

Method

Following the procedures in Wong, Schwartz and Jenkins (2005), monosyllabic tone productions were collected from three-year-old Mandarin-speaking children in Taiwan and low-pass filtered to eliminate lexical information but retain tone information. Five Taiwan adults categorized these filtered tones and those produced by the Mandarin-speaking children growing up in the U.S. reported in Wong, et al., (2005). Agreements on tone categorization by judges residing in Taiwan and in the U.S. were evaluated. Tone accuracy of children growing up in Taiwan and the U.S. were examined and compared.

Results

The U.S. and Taiwan judges showed high agreements on tone categorization. None of the four tones produced by the U.S. or Taiwan children was adult-like. Taiwan children made more errors in Tone 2 and Tone 4 than Mandarin-speaking children growing up in the U.S. Accuracy rates of Tone 1 and Tone 3 were comparable in the two groups of children.

Conclusion

Mandarin tone acquisition is a protracted process. Three-year-old Mandarin-speaking children growing up in Taiwan and the U.S. show similar developmental patterns and have not yet produced monosyllabic tones with adult-like accuracy.

Lexical tone (hereafter tone), the use of fundamental frequency (F0) or pitch as a primary means of contrasting lexical meanings, is an important feature in most of the languages in the world (Yip, 2002). Previous studies that examined children's acquisition of Mandarin tones employed different methodologies and found discrepant results. For example, some studies examined Mandarin tone acquisition in children growing up in an English-speaking environment, while others examined children growing up in a Mandarin-dominant environment; some studies determined children's tone accuracy by the judgments of Mandarin speakers residing in the U.S.; whereas some employed judges residing in a Mandarin-speaking environment. The age of acquisition of Mandarin tones reported in the studies ranged from less than two years to over six years of age and the order of acquisition of the tones were inconsistent across studies. Thus, the developmental course of Mandarin tone acquisition remains unclear. This study examined production of monosyllabic lexical tones by three-year-old Mandarin-speaking children growing up in Taiwan and compared their production accuracy to that of Mandarin-speaking children growing up in the U.S. to determine how well three-year-old children master the production of the four Mandarin tones in isolated monosyllabic words. To facilitate future cross-study comparisons, agreements on tone judgment by Mandarin-speaking adults residing in the U.S. and in Taiwan were compared.

Mandarin, the most widely spoken tone language, is a dominant and an official language in China and Taiwan. Each Mandarin syllable carries one of the four full tones -- Tone 1 (T1), Tone 2 (T2), Tone 3 (T3), Tone 4 (T4) -- or a neutral tone. When produced in isolation, the four tones have a high level, low rising, dipping (low falling-rising), and high falling pitch/F0 contour, respectively (See Figure 2 in Xu, 1997). T3 undergoes contextual variations. It is a low falling tone in non-final position, and a low falling or low dipping tone in utterance final position

(Duanmu, 2007). Though most speakers produce T3 in isolation with a low dipping F0 contour, some speakers, particularly Mandarin speakers in Taiwan, produce it as a low falling tone. According to the Target Approximation Model of Tonal Contour Formation (Xu & Xu, 2004; Xu, 2001a; Xu, 2001b), the pitch target for the four tones are high, rise, low and fall, respectively, suggesting that the articulatory goal for T3 is to reach a low F0. For the tones to be accurately perceived, the F0 contour is the primary and sufficient acoustic cue (Fu & Zeng, 2000; Massaro, Cohen, & Tseng, 1985). Though the amplitude contour and syllable duration covary with the four tones, they are negligible when F0 information is available (Fu & Zeng, 2000). Thus, to correctly produce the four tones, children need to master the production of the corresponding F0 contours. The neutral tone occurs in weakly stressed syllables and does not occur in monosyllabic words in Mandarin. It was, therefore, not included in the present study.

Large discrepancies have been found in children's age of acquisition of Mandarin tones. A longitudinal study involving four children (Hua, 2002) and a large–scale cross-sectional study involving 129 children (Hua & Dodd, 2000) carried out in Beijing, China, reported that the children produced the four tones accurately in various contexts before the age of two years. For example, Hua and Dodd (2000) found only two tone errors in all the productions of the 129 children, including 21 children between the age of 1;6-2;0, in a picture naming task and a picture description task, suggesting essentially no tone errors in children as young as one-and-a-half years old. The case study presented by Chao (1973/1951) on a Mandarin-speaking child growing up in the U.S. also suggested early tone acquisition. The child in the study reportedly produced the four full tones correctly at two years and four months old (Chao, 1973/1951).

Yet, Li and Thompson (1977) and Clumeck (1977, 1980) suggested that children do not acquire the four tones before the age of three years. Clumeck (1980) performed a longitudinal study on the development of Mandarin lexical tones in three Mandarin-speaking children growing up in the U.S. Two of them had not mastered the four tones by the end of the study at 2;10 and 3;5 (Clumeck, 1977; Clumeck, 1980). Li and Thompson (1977) collected tones produced in spontaneous speech and in a picture naming task over a seven-month period from 17 children between the ages of 18 to 36 months growing up in Taipei, Taiwan. The age of acquisition of tones was not specified but the study reported that the children did not produce the four tones correctly until they combined more than 2-3 words in their utterances (Li & Thompson, 1977).

More recently, Wong and colleagues reported several studies on the production of Mandarin tones by children learning Mandarin as their first language in the U.S. and found much more protracted course of development in the acquisition of the four tones. Wong, Schwartz and Jenkins (2005) collected monosyllabic tone production data from four adults and 13 three-year-old children using a picture naming task. The tones collected were low-pass filtered to eliminate segmental information but retain F0 information. Ten judges residing in the U.S. were recruited to evaluate the tones based on the filtered stimuli. The tones of the control (adult) group were easily identified, as expected. However, the children's tones differed from the adults' tones in key ways. Wong (2012) performed acoustic analysis on the tones produced by the adults and children reported in Wong et al., (2005), and confirmed that the three-year-old children's monosyllabic tones were phonetically different from the adults'. The children's T1 was not as high or as level as those of the adults. The children had more difficulties producing low frequencies: their T2 did not start as low as the adults', most of the children did not reach the low F0 target for T3, and the children's T4 ended at a higher F0 than the adults'. Given that previous EMG (electromyographic) studies showed involvement of the Sternohyoid muscle in producing these low F0 targets during Mandarin tone production (Erickson, Baer, & Harris, 1983; Hallé, 1994; Sagart, Halle, Boysson-Bardies, & Arabia-Guidet, 1986), the author suggested that children have more difficulties controlling a less frequently used laryngeal muscle for tone production and the development of tone production is limited by physiological constraints and maturation of speech motor control.

Wong (2008) examined disyllabic tones produced by 12 adults and 44 children growing up in the U.S. using the same procedures in Wong et al., (2005) and found that even 5- and 6-year-old children did not produce disyllabic tones with adult-like accuracy. Children's accuracy of the same tone varied depending on context. Disyllabic tone combinations with more complex F0 contours were more difficult for children to produce. For example, despite the fact that both T1-T2 (T1 followed by T2) and T2-T1 (T2 followed by T1) are a combination of T1 and T2, children had more difficulties producing T1-T2 (47% accuracy) than T2-T1 (81% accuracy) because T1-T2 has a more complex F0 than T2-T1. In T1-T2, the first tone, T1, is a high level tone ending in a high F0, while the second tone, T2, is a rising tone starting at a low F0. When producing T1-T2, a large transition is required to go from the high F0 at the offset of the first syllable to the low F0 onset for the tone in the second syllable, resulting in a high level, followed by a high falling, and then by a low rising F0 contour in the disyllabic tone. On the other hand, when producing T2-T1, the F0 for T2 ends at a high F0, close to the high F0 onset for T1 in the second syllable. Thus, the resulting F0 contour for the disyllabic tone is a high level F0 contour followed by a high falling F0 contour, less complex than the F0 contour in T1-T2. As a result, the author suggested that tones with more articulatory motor demands are more difficult for children to produce, supporting the physiological account of tone development suggested by Wong (2012).

Several methodological differences could have contributed to such large discrepancies in the age of acquisition of Mandarin tones in previous studies. Early studies employed only one judge, who was usually the experimenter, to determine tone accuracy by listening to the naturally produced (i.e., unfiltered) tones with contextual information. With lexical, segmental and/or contextual information available, it is difficult to control for lexical expectation in tone judgment. For example, when a child looks at a ball in the room and vocalizes, one would expect and tend to hear that the child was saying the word ‘ball’. Lexical knowledge and expectation have been found to influence the listener’s ability to detect differences in speech sounds and cause perceptual illusions (Oller & Eilers, 1975; Kent, 1996). Thus, lexical expectation of the experimenters in the earlier studies may have led to the findings of early tone acquisition in children.

Determining children’s tone accuracy based on productions in various tonal contexts could be another contributing factor to the divergent findings. All early studies collected tone productions from contexts ranging from monosyllabic to multisyllabic words and from isolated syllables to continuous speech. Tone accuracy was determined by collapsing the tones across all tonal contexts. Given that the F0 contour of the same tone varies depending on contexts (Xu, 1997; Xu, 2001b), children’s different performance in different studies could have been an effect of different tonal contexts of the productions.

The use of different criteria to determine acquisition may have also led to different results in age of acquisition of tones. Most earlier studies did not specify how tone mastery was defined and did not provide accuracy scores. Hua and Dodd (2000) and Hua (2002) defined tone stabilization as when 90% of the children in a specific age group produced the tone with at least 66.7% accuracy.

The three studies by Wong and colleagues were the only studies that controlled for lexical expectation of the judges by using low-pass filtering to eliminate segmental information, controlled for context effects by examining tones produced in isolated monosyllabic and disyllabic words only, and employed multiple judges who were blind to the design of the study to determine the tones produced by the children. These studies were also the only studies that determined children's tone mastery by comparing children's tone productions to adults'. However, given that these studies examined tone productions by Mandarin-speaking children growing up in the U.S. and employed Mandarin-speaking judges residing in the U.S. to categorize the tones, it is unclear whether the findings of late tone acquisition was due to the lack of Mandarin input and support to the children in the ambient English-speaking environment and whether the reported developmental pattern can be generalized to children growing up in places where Mandarin is a dominant language in the ambient environment such as Taiwan and China. In addition, the judges in the study were doctoral students residing in the U.S. It is unclear whether their tone judgment was influenced by English exposure and whether their tone judgments were different from or as reliable as the tone judgments by judges residing in Mandarin-speaking environments.

No study has compared Mandarin lexical tone productions in children growing up in different language environments. Given that bilingualism is the norm in the world, similarities and differences in the developmental patterns in children growing up in different language environments will provide insights into universal factors that affect tone development and information on the effect of environmental input. This information will be valuable for understanding phonological development and for educators, speech language pathologists and linguists who work with Mandarin-speaking children learning their first language in different language environments.

To validate the design and findings of previous research that reported protracted tone development, such as those conducted by Wong and her colleagues, and to lay the ground work for future cross study comparisons of tone development in children growing up in different linguistic environments, it is necessary to ensure that judges recruited in different linguistic environments such as Taiwan and the U.S. are equally reliable in their tone judgment and to determine factors that may affect interjudge reliability among judges from different linguistic environments.

The first goal of the present study was to determine the degree of agreement and inter-judge reliability on tone judgments among adults recruited in Taiwan and in the U.S. and to explore factors that affect interjudge reliability on tone judgments. The second goal of the study was to examine the similarities and differences in tone accuracy, defined as the perceptual accuracy of the target tones by native listeners, in monosyllabic words in three-year-old children growing up in Taiwan and the U.S. To achieve these goals, monosyllabic tone productions by 11 children growing up in Taiwan were collected. Tone accuracy was judged by five adults in Taiwan. The same five adults also re-categorized the tones produced by the four adults and the 13 children growing up in the U.S. reported in Wong et al., 2005. Tone judgments on the tones produced by the four adults and 13 children in the U.S. reported in Wong et al., (2005) by the five Taiwan judges and the 10 U.S. judges employed in Wong et al., (2005) were compared for interjudge reliability. Results of tone production accuracy of children growing up in the U.S. reported in Wong et al., 2005 were re-examined using the tone judgments by the Taiwan judges to determine if there was any major difference in the findings. The accuracy rates of tones produced by the U.S. and Taiwan children based on the judgment of the same five Taiwan judges were compared to examine the effect of ambient language environment on children's acquisition of lexical tones.

The specific research questions are: (1) Is there any significant difference on tone judgments by Mandarin-speaking adults residing in the U.S vs. in Taiwan? (2) What factors affect inter-judge reliability of tones? (3) How well do children growing up in the U.S. produce the four tones based on the tone judgment of Taiwan judges? (4) How well do children in Taiwan master the production of Mandarin tones in monosyllabic words based on the tone categorization of the Taiwan judges? (5) What are the similarities and differences in the tone production accuracy rates of children learning Mandarin as their first language in Taiwan and in the U.S.?

Method

This study followed the procedures used in Wong, et al., (2005).

Speakers

Three-year-old Children in Taiwan (Taiwan children). Eleven three-year-old children (6M, 5F, mean age=3;1, age range = 2;10-3;6) were recruited in Chiayi, Taiwan. Eight of them were attending pre-school while three had not started school. All children were administered a hearing screening, and a Mandarin speech and language test-- Language Disorder Scale of Preschoolers (學前兒童語言障礙評量表, LDSP) (Lin & Lin, 1994). A language sample was collected and parents of the children provided information on the children's language background and developmental history.

Though Mandarin is the dominant language in Taiwan, Taiwanese (Southern Min dialect) is extensively used in most parts of Taiwan, except for the metropolitan areas in Taipei. Therefore, additional language criteria involving Mandarin proficiency, amount of Mandarin exposure and amount of Mandarin use were adopted to ensure that the children were acquiring Mandarin as their first language in Taiwan. All children included in the study met the following inclusion criteria: (1) Mandarin was the dominant language in the home, (2) parents reported that the child had limited exposure to Taiwanese and was exposed to Mandarin over 80% of the time and spoke Mandarin over 90% of the time during the day, (3) if the child had started schooling, Mandarin was the dominant language and language of instruction in the child's school, (4) the child obtained a score above the 16th percentile in LDSP , a Mandarin speech and language test normed in Taiwan, (5) the child passed the hearing screening, (6) there was no limitation or atypicality observed in the Mandarin language sample collected, and (7) the child had unremarkable motor, social, emotional, cognitive and language development per parent report. Table 1 lists the demographic information and language background of the children.

Three-year-old Children in the U.S (U.S. children). The tone productions of the 13 children reported in Wong et al., 2005 (7M, 6F, mean age = 3;0, age range = 2;10-3;4) being judged by Mandarin-speakers residing in the U.S. were reassessed in this study by asking the judges recruited in Taiwan to re-categorize the tones. As reported in Wong et al., 2005, the 13 children were monolingual Mandarin-speaking children growing up in the U.S. learning Mandarin as their first language. The children were exposed to Mandarin at home, obtained a total language score above 16th percentile in LDSP, passed the hearing screening, did not demonstrate any language limitation in the language sample, and had unremarkable developmental history. All children had limited exposure to English and none of them went to

English day care centers. The Preschool Language Scale—3 (Zimmerman, Steiner, & Pond, 2002) was administered to ensure the monolingual status of the children. All children scored more than one Standard Deviation below the mean in the total score in PLS-3 (mean percentile rank = 2%. See Wong et al., 2005 Table 2 for details).

Adults (U.S. adults). Adult productions were the tones produced by the four mothers reported in Wong et al., (2005). Three of the adults were from China and one was from Taiwan. They all reported to be dominant in Mandarin and continue to use Mandarin in the U.S.

Stimuli for Tone Production

The same 24 pictures employed in Wong et al., (2005) depicting 24 monosyllabic words were used to elicit tone productions from the Taiwan children. Half of the words formed six minimal pairs contrasting the four tones and the other half were singletons with no minimal contrast in tone. See the Appendix in Wong et al., (2005) for the word list.

Tone Collection Procedures

Tone collection was conducted in the child's school or in a quiet room in National Chung Cheng University in Taiwan. Tone productions were collected by the same experimenter using the same procedures in Wong, et al. (2005). The 24 pictures were presented to the children one by one. Questions such as "這是什麼 [What is this]"or "他在幹嗎? [What is he doing]?" were used to elicit tone productions. If the child did not produce the target word, a toy object or a real object was presented. If the child continued not to produce the target word, semantic prompts were given. Each picture was labeled two times. After the picture naming task, the Mandarin test was administered. Then a Mandarin language sample was collected and the hearing screening was administered. Children's productions were recorded on a digital recorder using a dynamic microphone.

Taiwan Judges

Five students (3F, 2M, 4 undergraduates and 1 graduate student) at National Chung Cheng University in Taiwan were recruited as judges to listen to the tones produced by the Taiwan and U.S. children and adults. Their mean age was 20 years (range = 18;3-23;9). All of them reported that Mandarin was their dominant and strongest language. None of the judges had visited or resided in other countries. All judges, except one (TJ05) learned Mandarin since birth. TJ05 learned Mandarin at seven years of age. Four of the judges (TJ02-TJ05) reported Taiwanese and English to be their second and third languages, respectively. One judge (TJ01) indicated English as her second language and Taiwanese as her third language. TJ02, TJ03 and TJ04 reportedly used Mandarin almost exclusively (95% to 100% daily). TJ01 used Mandarin, Taiwanese, English and Hakka 80%, 10%, 5% and 5% of the time, respectively, in a day. TJ05 used Mandarin 70%, Taiwanese 20% and English 10% of the time in a day. See Table 2 for information on the Taiwan and U.S. judges.

Stimuli for Tone Judgment

Experimental Stimuli. The words produced by the U.S. and Taiwan children and adults were presented as stimuli in the experimental blocks for the five Taiwan judges to categorize the tones. These included the 198 child productions and the 92 adult productions collected in the U.S. reported in Wong et al., (2005), and the monosyllabic target words spontaneously produced in isolation by the eleven Taiwan children described above. Following the procedures in Wong et al., (2005), non-target words, noisy productions (i.e., productions with overlapping of voices, loud background noise, or interruptive clicks and pops), clipped productions, unintelligible productions (e.g., mumbles and very soft productions) and productions that were not produced in isolation were excluded. Because the word “mao4” was excluded in Wong et. al., (2005) due to the fact that none of the U.S. children produced the word in isolation, this word was also excluded for analyses in the Taiwan children's productions. Altogether the Taiwan children produced 170 usable tone productions (46, 42, 33, 49 for T1, T2, T3, and T4, respectively).

Following the procedures in Wong et al., (2005), low-pass filtering was used to eliminate the lexical information and retain F0 information in the words produced by the U.S. and Taiwan children and adults before presenting the tones to the judges for tone judgment. Adult productions were low-pass filtered at 400 Hz while child productions were low-pass filtered at 500 Hz. All productions were normalized for intensity and were blocked by speakers. Thus, there were four blocks of adult stimuli and 24 blocks of child stimuli (13 blocks from the U.S. children and 11 blocks from Taiwan children).

Training Stimuli. The same 48 training stimuli (4 tones x 12 monosyllables) used in Wong et al., (2005) were used for the training blocks in this study to familiarize the Taiwan judges with the tone categorization task. All the training stimuli were naturally produced (i.e., unfiltered) by a female Mandarin speaker. Twenty-four of the stimuli were morphemes in Mandarin, whereas the other 24 stimuli were nonsense syllables with legal segmental construction in Mandarin. The 48 stimuli were put in two blocks. Each block consisted of 12 Mandarin morphemes (3 morphemes x 4 tones) and 12 nonsense syllables (3 syllables x 4 tones).

Tone judgment procedures

Tone judgment procedures were similar to those in Wong et al. (2005). In addition to the tones produced by the Taiwan and U.S. children and adults, the five Taiwan judges also categorized additional monosyllabic tone productions for a larger study. Each judge attended two 30-45 minute sessions every day for five days in a quiet room. The two sessions on the same day were separated by at least one hour. In the first session, the judges had a hearing screening, filled out a language background questionnaire, and listened to a block of training stimuli. All judges attained over 90% accuracy in the training block (mean=95%, range=92%-100%). Then the judges listened to nine blocks of experimental stimuli, with a total of 151 trials. In the subsequent sessions, the judges were randomly presented with 5-7 blocks of stimuli with a total of 166-178 trials. In the 3rd, 5th and 9th session, the judges were randomly presented one of the two blocks of training stimuli to monitor their attention level. The judges attained a mean accuracy of 91% in the training blocks across all sessions (range = 85%-94%). In the 7th session, the judges redid one of the experimental sets for determining intrajudge reliability.

A customized computer program (Tagliaferri, 2005) was used to present the stimuli for tone categorization. When the judge was ready, s/he clicked the start button on the screen. The computer program randomly picked a block from the experimental set and randomly presented one of the sounds in the block. The judge listened to the sounds at a comfortable level via headphones. Because determining tones in filtered speech without lexical support was a more challenging task, the judges were allowed to click the replay button to listen to the sound as many times as s/he wanted. After s/he made a decision by clicking one of the four response buttons indicating Tone 1, Tone 2, Tone 3, and Tone 4 on the screen, the next trial began. After a block of trials was presented, the judge could take a break or click the “start” button to continue to the next block.

Data Analysis

Interjudge Reliability of the U.S. and Taiwan Judges

To answer the first question on the agreement on tone categorization by the Taiwan and U.S. judges, group and individual inter-judge reliability among the judges on their categorization of the tones produced by the U.S. children and adults was examined. Fleiss’s kappa coefficient was used to measure the inter-judge agreements of the three groups of judges -- the 10 U.S. judges reported in Wong, et al., 2005, the five Taiwan judges described above, and all the 15 judges from the U.S. and Taiwan -- on their categorizations of the tones produced by the 13 U.S. children and 4 adults. The number of T1, T2, T3 and T4 responses by the groups of judges on each of the 290 productions by the 13 U.S. children and 4 adults were tallied and used as the sampling variable. Table 3 shows the Fleiss's kappas for the three groups of judges. Fleiss' kappa coefficients less than 0 indicate no agreement while coefficients of 0-.20, .21-.40, .41-.60, .61-.80, and .81-1 indicate slight, fair, moderate, substantial and almost perfect agreement, respectively. As shown, the three groups of judges showed similar reliability patterns. Substantial agreement was found within the three groups of judges on their categorization of T1, T2, T4 and across the four tones produced by the U.S. children and adults (Table 3 top cluster). Moderate agreement among the three groups of judges were reached for T3 productions. Overall, the U.S. and Taiwan judges showed substantial reliability on their tone judgments.

Pairwise comparisons were performed using Cohen's kappa and percent agreement to examine the reliability of each judge to every other judge on their categorization of tones produced by the U.S. children and adults. Cohen's kappas were computed on the number of T1, T2, T3, and T4 responses by each pair of judges on their categorization of each of the 290 tones produced by the U.S. adults and children. Percent agreement was defined as the percent of the times the two judges categorized the 290 productions into the same tone category regardless of the intended tone of the speaker. Table 4 shows the results. The upper right half of the table shows the Cohen's kappa coefficients between each pair of judges. Like Fleiss' kappa, Cohen's kappa coefficients of 0-.2, .21-.4, .41-.6, .61-.8, and .81-1 indicate slight, fair, moderate, substantial and almost perfect agreement, respectively. As shown, of the 120 pair-wise comparisons, 90% (n=108) reached substantial agreement, 0.02% (n=2) reached almost perfect agreement, and 10% (n=10) reached moderate agreement, indicating high individual reliability of the judges. Nine of the 10 instances of moderate agreement involved the categorizations performed by the Taiwan judge TJ05. Similar patterns were observed based on the percent of agreement, indicating high interjudge reliability among the 15 judges. Taken together, the Taiwan and U.S. judges demonstrated high agreement and high group and individual reliability in their categorization of the tones produced by the children and adults.

Factors that Influence Interjudge Reliability

To answer the second question on factors that influence inter-judge reliability on tone judgment, interjudge reliability of judges recruited in Taiwan vs. in the U.S. and judges who originally came from Taiwan vs. from China were examined for systematic differences. Moreover, accuracy rates of the four tones produced by the U.S. speakers determined by the Taiwan judges and by the U.S. judges were compared for differences. Furthermore, Pearson correlation coefficient was used to determine if there was any association between production accuracy and interjudge reliability.

From the reliability analyses above, it appears that inter-judge reliability was related to degree of accuracy of the tones. As indicated in Table 3, Fleiss' kappas were the highest on the tones produced by the adults and lowest on children's productions, and among the four tones, T3 productions, which have been reported to be easily confused with T2 in previous literature on adults' perception of Mandarin tones (Shen & Lin, 1991), also yielded lower kappa coefficients than the other tones. Pairwise comparisons on the agreement of judges on the children's and adults' productions using Cohen's kappas also showed similar patterns. Cohen's kappa coefficients among the 15 judges ranged from .72 to .97 for adults' productions, with 87% (N=108) of the 120 comparisons yielded almost perfect agreement and 13% (N=16) reached substantial agreement. Inter-judge reliability among the 15 judges was lower for the tones produced by U.S. children. Cohen's kappa coefficients ranged from .44 to .77 with 63% of the comparisons (N=76) yielded substantial agreement and 37% (N=44) yielded moderate agreement. Pearson correlation coefficient showed almost perfect correlation between Fleiss's kappas and percent correct of the four tones produced by U.S. children and adults and judged by the 15 Taiwan and U.S. judges, r (N = 8) = .96; p < .001, r^2 = .93. Pearson correlation coefficient between Fleiss's kappas of the five Taiwan judges and the judged accuracy rates of the tones produced by the U.S. and Taiwan speakers based on the judgments of the five Taiwan judges also showed very high correlation, r(N = 12) = .93, p < .001, r^2 = .87.

Patterns in the Cohen's kappa coefficients in Table 4 were examined to see if there was any systematic difference in the reliability between judges who were residing in the U.S. vs. in Taiwan. Because five of the U.S. judges originally came from Taiwan to the U.S. (Tables 2 & 4), patterns of reliability between judges from the same vs. different country of origin were also examined. As indicated in Table 4, the judges all achieved similar degrees of agreement with other judges regardless of the countries they were residing in or their country of origin. For example, TJ05 had the lowest interjudge reliability with other judges. The judges that had the highest and the lowest agreements with TJ05 included judges from both Taiwan and China. Similar patterns were also found with the judges' reliability on adult productions only and child productions only.

In all, interjudge reliability on tone categorization was affected by degree of accuracy of the tone productions. Judges had higher interjudge reliability when categorizing adults' tones, which were more accurately produced, than children's tones. Interjudge reliability was not affected by the country of origin or the country of residence of the judges. No systematic difference was found in the reliability of judges from Taiwan or China or judges residing in Taiwan or the U.S.

Tone Accuracy of U.S. Children Determined by Taiwan Judges

To answer the third question on tone production accuracy, defined as perceptual accuracy of the target tone by native listeners, of U.S. children, and to validate whether the findings reported in Wong et al., (2005) on U.S. children's tone productions still held if the tones were judged by adults in Taiwan rather than adults residing in the U.S., tone accuracy of the U.S. children and adults determined by the Taiwan judges were examined. The five Taiwan judges categorized the four tones of U.S. children with 67%, 67%, 47%, and 79% accuracy (Table 5a, left panel) and the four tones of U.S. adults were categorized with 93%, 96%, 82%, and 94% accuracy (Table 5b, left pane). None of the four tones produced by the U.S. children were judged by the Taiwan judges as having achieved adult-like accuracy, χ^2(N=395, df=1) = 31.06; p < .001, w = .3 for T1, χ^2(N=390, df=1) = 36.97; p < .000, w = .3 for T2, χ^2(N=355, df=1) = 38.90; p < .001, w = .3 for T3, and χ^2(N=310, df=1) = 11.73 ; p = .001, w = .2 for T4. A Chi-square test revealed that the accuracy rates for U.S. children's T4 was better than the other three tones (Table 6). T3 was the worst. No difference was found between the accuracy rates on U.S. children's T1 and T2. Overall, the order of the accuracy rates of the four tones by the U.S. children judged by the Taiwan judges was T4 > T1 = T2 > T3 (Table 6). The error patterns in Table 5a show that U.S. children's T1 was mostly categorized as T2 or T4. The judges confused U.S. children's T2 and T3 productions, with most T2 errors being categorized as T3 and T3 errors being categorized as T2. Some T3 errors were misperceived as T4. The U.S. children's T4 errors were mostly perceived as T3 (Table 5a, left panel and Table 7).

Little difference was found in the results based on the Taiwan judges reported in this study vs. the results based on the U.S. judges reported in Wong et al., (2005). Neither group of judges categorized any of the tones produced by U.S. children with adult-like accuracy. The accuracy rates and error patterns of the tones of the U.S. adults and children judged by the two groups of judges were strikingly alike (Tables 5a and 5b), except for T1 produced by the U.S. children. The Taiwan judges identified more errors in the U.S. children's T1 productions than the U.S. judges, χ^2(N=825, df=1) = 11.81; p < .001, w = .12. The order of accuracy determined by the Taiwan judges was T4 > T1 = T2 > T3, Table 6), which is only slightly different from the order of accuracy determined by the U.S. judges (T4 = T1 = T2 > T3) reported in Wong et al., (2005).

In sum, tone judgments performed by the Taiwan judges on the tones produced by the U.S. speakers by and large replicated the findings in Wong et al., (2005) and confirmed the findings reported in Wong et al., (2005) that children growing up in the U.S. have not acquired the production of the four tones and have more difficulties producing T3.

Tone accuracy of Taiwan Children Judged by Taiwan Judges

To answer the fourth question on tone accuracy of Taiwan children, first, intra-judge reliability of the five judges on Taiwan children's productions was examined. Then, the Taiwan children's tone accuracy categorized by the Taiwan judges was determined and compared to those of adults to determine mastery. Error patterns on children's productions were also examined.

Intrajudge reliability of the Taiwan judges on the tones produced by Taiwan speakers all reached substantial to almost perfect agreement. Cohen's kappa coefficients ranged from 0.63 to 0.82 and the percent of agreement between the judges' first and second rating of the same tones ranged from 73% to 87%.

Taiwan children's tones were identified by the Taiwan judges with 71%, 53%, 53%, and 68% accuracy, respectively (Table 5c). None of Taiwan children's tones were adult-like, χ^2(N=350, df=1)=22.95, p <.001, w=.3 for T1; χ^2(N=330, df=1)=65.35, p < .001, w=.4 for T2; χ^2(N=285, df=1) = 25.55, p < .001, w = .3 for T3; χ^2(N=345, df=1) = 25.74, p < .001, w = .3 for T4. Taiwan children's T1 errors were mostly categorized as T2. Their T2 errors were mostly perceived as T3 and sometimes as T1. Their incorrect T3 productions were mostly perceived as T2 and sometimes as T4. Taiwan children's T4 errors were categorized by the Taiwan judges as T3 or T1 (Tables 5c & 7). The Pearson Chi-square test revealed that the accuracy rates of Taiwan children's T1 and T4 were better than T2 and T3 (Table 6). No significant difference was found in their accuracy rates between T1 and T4 or T2 and T3. Thus, the order of accuracy of the tones produced by Taiwan children from the highest to the lowest was: T1, T4 > T2, T3.

Comparing Tone Accuracy between U.S. and Taiwan Children Based on the Judgments of Taiwan Judges

To answer the fifth question on the similarities and differences in tone accuracy between children growing up in the U.S. and in Taiwan, tone judgments of the Taiwan judges on the U.S. and Taiwan children were compared. First, analyses were performed to determine whether the two groups of children differed in their Mandarin language scores. Because the data did not violate the assumptions for parametric statistics, t-tests for independent samples were used to compare the receptive, expressive and total scores in the U.S. and. Taiwan children measured by the Language Disorder Scale of Preschoolers. The results showed non-significant difference in all the comparisons, indicating that the two groups did not differ in their Mandarin language scores as measured by the Language Disorder Scale of Preschoolers.

Next, we compared the accuracy rates of the U.S. and Taiwan children judged by the Taiwan judges. As indicated above, neither the children in the U.S. or in Taiwan produced any of the four tones with adult-like accuracy. As shown in the left panels of Tables 5a and 5c, similar error patterns were found in the Taiwan judges' categorization of the tones produced by the U.S. and the Taiwan children. The similarity and differences in the error patterns were summarized in Table 7. Neither group of children produced any of the tones with adult-like accuracy. T1 errors were mostly perceived as T2, T2 as T3, T3 as T2 or T4 and T4 as T3. Pearson's Chi-Square test was used to examine whether there was any difference in the accuracy rates of the four tones produced by the Taiwan and the U.S. children based on the judgments by the five Taiwan judges. The results showed that the Taiwan and U.S. children produced T1 and T3 with similar accuracy, but Taiwan three-year-old children produced T2 and T4 with significantly lower accuracy than the three-year-old children growing up in the U.S., χ^2(N=480, df=1)=10.52, p=.001, w=.1 for T2; χ^2(N=455, df=1)=6.21, p=.013, w=.1 for T4, which explains the difference in the order of tone accuracy in the two groups of children (T4 > T1 = T2 > T3 vs. T4 = T1 > T2 = T3, Table 7).

To summarize, the tone accuracy rates and error patterns of the U.S. and Taiwan children were by and large similar. Neither Taiwan nor U.S. children produced adult-like tones, and T3 was more difficult than T1 or T4 for both groups. Taiwan children made more errors than U.S,. children with T2 and T4, resulting in comparable accuracy rates between T1 and T4 and between

T2 and T3 in Taiwan children's tones. Thus, the order of accuracy of the four tones was slightly different in the two groups of children.

Discussion

The first goal of the study was to determine the agreement of tone categorization by Taiwan and U.S. adults and to explore factors that may influence interjudge reliability of tone categorization. The results showed high interjudge agreement among all judges regardless of their location and country of origin. The 15 Taiwan and U.S. judges as a group showed almost perfect agreement on their categorization of adults' tones and fair to substantial agreement on categorizing children's tones (Table 3). Interjudge reliability among the ten U.S. judges was as high as among the five Taiwan judges (Table 3). Each of the 15 U.S. and Taiwan judges showed similar patterns of agreement with the other 14 judges on their tone categorization (Table 4). Exclusion of any of the judges had little effect on the overall interjudge reliability.

The almost identical accuracy scores and error patterns in the categorization of adults' tones by the U.S. and Taiwan judges (Table 5b) and the similar results between the two groups of judges on their categorization of the tones produced by the U.S. children (Table 5a) provided additional evidence to support the claim that judges residing in the U.S. and in Taiwan and judges who were brought up in Taiwan and in China do not differ in their tone judgments. The only observed difference in the tone categorization by the U.S. and Taiwan judges was that Taiwan judges made more errors on categorizing U.S. children's T1 productions (Table 5a). This is unlikely to be due to differences in T1 perception between the Taiwan and U.S. judges, given that the two groups of judges did not differ significantly in the judgments of the adults' T1. Acoustic data on the tones produced by the U.S. children reported in Wong (2012) showed that even when the target tone was correctly identified by the judges, the fundamental frequency contours of children's T1 productions were not as level nor as high as those of adults', suggesting that even children's correctly perceived tones included intermediate productions that were phonetically not fully target-like. Thus, the differences in Taiwan judges' categorization of U.S. children's T1 productions may be attributed to differences in the criteria in determining T1 productions when the high and level pitch targets for T1 were not fully achieved. One possible reason for this difference between the U.S. and Taiwan judges could be because Taiwan judges were exposed to Taiwanese which has two level tones, a high level and a mid level tone, causing the Taiwan judges to have a slightly different phonetic category boundary for T1 in terms of F0 height and F0 shifts.

Findings of the current study suggest that interjudge reliability is associated with degree of accuracy of the tones produced. The judges reached much higher agreement in their categorization of adults' tones than children's tones (Table 3). Among the four tones, the least amount of agreement was found in the judges' categorization of T3 (Table 3), which has been found to be the most difficult to categorize even in adults' speech (Gandour, 1978; Shen & Lin, 1991; Whalen & Xu, 1992). The strong positive correlations between Fleiss' kappas and judged accuracy of the adults' and children's four tones provided further support to this claim. This high correlation is not surprising considering that productions that are produced with the exact target form are not ambiguous, and should elicit the same tone categorization from different judges. On the other hand, incorrectly produced tones that do not match any of the tone categories, imprecise productions that undershoot or overshoot the tonal targets, and intermediate productions that partly resemble more than one tone categories will be more ambiguous and, thus, should be less likely to elicit consistent categorizations among the judges.

These findings on interjudge reliability have implications for future studies on tone acquisition. First, the number of judges required in future studies can be significantly reduced given the high group and individual reliability among the judges. Second, the high interjudge reliability between Taiwan and U.S. judges suggests that tone judgments by native Mandarin speakers recruited in Mandarin-dominant and English-dominant environments do not differ significantly. Thus, it implies that it is less important to control for country of residence than to ensure that a) Mandarin is the dominant and family language of the judges, b) the judges started to learn Mandarin before 3 years of age (TJ05 started learning Mandarin at seven years of age and demonstrated relatively lower tone judgment agreement with other judges) in their native country, and c) the judges continue to use Mandarin in their daily life. The high reliability between the Taiwan and U.S. judges also allows comparisons of findings across studies that employ native Mandarin-speaking judges from different language environments. Third, given the finding that interjudge reliability varies with phonemic accuracy, future developmental studies on phonemic acquisition may need to consider determining and reporting interjudge reliability on a control group that is expected to have high phonemic accuracy (e.g., an adult control group) in addition to reporting interjudge reliability of the experimental groups.

The second goal of the study was to compare the similarities and differences in monosyllabic tone accuracy in children growing up in Taiwan and the U.S. In terms of the tone accuracy by children growing up in the U.S., the findings in this study largely support the findings reported in Wong et al, (2005). Results of perceptual judgments by the U.S. and Taiwan judges consistently showed that three-year-old children growing up in the U.S. have not yet mastered the production of the four tones. None of their four tones was adult-like. T3 was the most difficult and was produced with the lowest accuracy rate. The major substitution patterns of children’s incorrect productions included T1 errors being perceived as T2, T2 errors being mostly mistaken as T3, T3 incorrect productions being largely perceived as T2 but sometimes as T4, and T4 errors being mostly categorized as T3 (Table 5). There is a slight difference in the order of accuracy of the four tones judged by the Taiwan vs. the U.S. judges due to the more T2 and T4 categorizations of the children’s T1 productions by the Taiwan judges. Both the U.S. and Taiwan judges indicated that U.S. children produced T1, T2 and T4 better than T3. The only difference was that the Taiwan judges judged that U.S. children’s T4 was significantly more accurate than T1 and T2, whereas the U.S. judges decided that U.S. children’s accuracy rates of T4 was similar to those of T1 and T2. These two seemingly different orders do not contradict the findings reported in Wong (2012) which examined the acoustic characteristic of the tones produced by these U.S. children. The author posited that perceptual judgment alone provides a broad index on children’s accuracy while the combination of perceptual judgments and acoustic data provides a more sensitive measure and reveals more fine-grained differences in children’s phonemic accuracy. Based on the perceptual and acoustic data, the study concluded that the order of acquisition of tones by the U.S. children from the most to the least accurate was T4 > T1 > T2 > T3.

Contrary to prediction, Taiwan children did not outperform U.S. children. Taiwan children’s accuracy rates on the four tones were 71%, 53%, 53%, and 68%, respectively (Table 5). None of the four tones was produced with adult-like accuracy. T2 and T3 were the most difficult and were produced with significantly lower accuracy rates than T1 and T4 (Table 6). The error patterns of Taiwan children’s tone productions were similar to those of the U.S. children except that there was seemingly fewer substitutions of T4 for T1 and more substitutions of T1 for T2 and T4 (Table 5).

Overall, the U.S. and Taiwan three-year-old children exhibited similar development in their production of monosyllabic Mandarin tones. Table 7 summarizes the similarities and differences in the accuracy rates of the tones produced by the Taiwan and U.S. children. Neither U.S. nor Taiwan three-year-old children learning Mandarin as their first language produced any of the tones with adult-like accuracy. T3 was particularly difficult for both groups of children. Both U.S. and Taiwan children did not differ in their accuracy rates for T1 and T3. Surprisingly, children in Taiwan produced T2 and T4 with lower accuracy rates than U.S. children, which led to a slightly different order of accuracy of the four tones in the two groups of children. Still, these orders do not contradict the order of accuracy of the four tones proposed in Wong (2012) using both acoustic and perceptual data, that is, T4 > T1 > T2 > T3.

It is surprising that children in Taiwan who presumably have more exposure to Mandarin than the U.S. children had lower accuracy rates in T2 and T4. Future studies that compare the acoustic characteristics of the tones produced by the U.S. and Taiwan children will provide more detailed differences in the tones produced by the two groups of children.

The by and large similar findings in the two groups of children growing up in different language environment may suggest possible biological universal constraints underlying the acquisition of Mandarin tones, which were proposed by previous research on Mandarin tone acquisition in children growing up in the U.S. Wong (2012) reported that three-year-old children did not produce the initial low fundamental frequency for T2, the low pitch target for T3 and the low fundamental frequency at the offset of T4 as low as adults. Previous electromyographic studies that examined the physiological mechanisms for tone production showed that these low fundamental frequency targets involve the use of the sternohyoid muscle and likely other strap muscles (Erickson, 1993; Hallé, 1994; Sagart et al., 1986), which are less frequently used in young children's vocalization and cries. This led the author to suggest that children's tone productions was limited by immature control of less frequently used laryngeal muscles.

Previous research also suggested that children's tone accuracy is related to articulatory complexity. Wong et al., (2005) and Wong (2012) reported that T3, which has the most complex fundamental frequency contour when produced in isolated syllables, was produced with the lowest accuracy rates by three-year-old children growing up in the U.S. and the fundamental frequency contours of children's T3 were the most varied and deviated the most from the target form. Comparing children's tone accuracy in different tonal contexts in disyllabic words, Wong (2008) reported that children are less able than adults to produce sequences of tones that have more complex F0 contours than tone sequences that have less complex fundamental frequency contours. The two- to six-year-old children in the study made significantly more errors in producing tone combinations in which the fundamental frequency offset in the first tone differ substantially from the onset fundamental frequency of the following tone (e.g., T4 followed by T4) than in producing tone combinations in which the offset frequency of the first tone is close to the onset frequency of the second tone (e.g., T4 followed by T3). Together with the finding that three-year-old children failed to maintain a consistently high and level frequency throughout a T1 syllable (Wong, 2012), the findings of these studies suggest that three-year-old children have not yet fully mastered the control and coordination of laryngeal muscles for tone production.

If children's tone production is limited by biological development and maturation of speech motor control, it is not surprising to find universal developmental patterns in tone acquisition in children learning Mandarin as a first language in different language environment with the condition that the children have received adequate language input and have been given sufficient opportunity to use the language in a natural environment. More research will be needed to confirm this hypothesis. If similar accuracy rates, substitution patterns, and order of acquisition of the tones are found in future studies that examine Mandarin tone development in children growing up in different language environments and if future acoustic studies find that Mandarin-speaking children growing up in different language environment consistently demonstrate difficulties producing tones with more complex F0 and producing very low pitch, more confident conclusions can be drawn.

Information on the universal developmental patterns in children growing up in different language environments will be particularly valuable for clinical assessment and treatment of phonological disorders. Nowadays given that bilingualism is the norm, most children are growing up in bilingual or multilingual environments in which the influence of another language or dialect is inevitable. It will be difficult, if not impossible, to establish developmental norms specifically for children coming from each unique linguistic environment. Thus, information on the general patterns of phonemic development in children growing up in different linguistic backgrounds can be used as a framework for speech language pathologists, educators and linguists working with children learning the same language in different parts of the world, such as Mandarin learning children in different parts of China and Taiwan, and immigrant children learning Mandarin in the U.S.

References

Chao, Y. R. (1973/1951). The Cantian idiolect: An analysis of the Chinese spoken by a twenty-eight-month-old child. In C. A. Ferguson, & D. I. Slobin (Eds.), *Studies of child language development* (pp. 13-33). New York: Holt, Rinehart & Winston.

Clumeck, H. (1977). Topics in the acquisition of Mandarin phonology: A case study. *Papers and Reports on Child Language Development, 14*(December), 37-73.

Clumeck, H. (1980). The acquisition of tone. In G. H. Yeni-Komshian, J. F. Kavanaugh & C. A. Ferguson (Eds.), *Child phonology: Vol. 1. production* (pp. 257-275). New York: Academic Press.

Duanmu, S. (2007). *The phonology of standard Chinese* (2nd ed.). New York: Oxford University Press.

Erickson, D., Baer, T., & Harris, K. S. (1983). The role of the strap muscles in pitch lowering. In D. M. Bless, & J. H. Abbs (Eds.), (pp. 279-285). San Diego, CA.: College-Hill Press.

Erickson, D. (1993). Laryngeal Muscle activity in connection with Thai tones. *Research Institute of Logopedics and Phoniatrics Annual Bulletin, 27*, 135-149.

Fu, Q. J., & Zeng, F. G. (2000). Identification of temporal envelope cues in Chinese tone recognition. *Asia Pacific Journal of Speech, Language and Hearing, 5*, 45-57.

Gandour, J. T. (1978). The perception of tone. In V. A. Fromkin (Ed.), *Tone: A linguistic survey* (pp. 71-76). New York: Academic Press.

Hallé, P. A. (1994). Evidence for tone-specific activity of the sternohyoid muscle in modern standard Chinese. *Language and Speech*, , 103-123.

Hua, Z. (2002). *Phonological development in specific context: Studies of Chinese-speaking children*. Clevedon, England: Multilingual Matters Limited.

Hua, Z., & Dodd, B. (2000). The phonological acquisition of Putonghua (modern standard Chinese). *Journal of Child Language, 27*(1), 3-42.

Kent, R. D. (1996). Hearing and believing: Some limits to the auditory-perceptual assessment of speech and voice disorders *American Journal of Speech-Language Pathology, 5*(3), 7-23.

Li, C. N., & Thompson, S. A. (1977). The acquisition of tone in Mandarin-speaking children. *Journal of Child Language, 4*(2), 185-199.

Lin, B., & Lin, N. (1994). *Language disorder scale of preschoolers(學前兒童語言障礙評量表).* Taipei: National Taiwan Normal University, Department of Special Education.

Massaro, D. W., Cohen, M. M., & Tseng, C. (1985). The evaluation and integration of pitch height and pitch contour in lexical tone perception in Mandarin Chinese. *Journal of Chinese Linguistics, 13,* 267-290.

Oller, D. K., & Eilers, R. E. (1975). Phonetic expectation and transcription validity. *Phonetica, 31*(3-4), 288-304.

Sagart, L., Halle, P., Boysson-Bardies, B. d., & Arabia-Guidet, C. (1986). Tone production in modern standard Chinese : An electromyographic investigation. *Cahiers De Linguistique - Asie Orientale, 15*(2), 205-221.

Shen, X. S., & Lin, M. (1991). A perceptual study of Mandarin tones 2 and 3. *Language and Speech, 34*(2), 145-156.

Tagliaferri, B. (2005). *Paradigm* [Perception Research Systems Inc.] Retrieved from www.perceptionresearchsystems.com

Whalen, D. H., & Xu, Y. (1992). Information for Mandarin tones in the amplitude contour and in brief segments. *Phonetica, 49*(1), 25-47. Wong, P., Schwartz, R. G., & Jenkins, J. J. (2005). Perception and production of lexical tones by 3-year-old, Mandarin-speaking children. *Journal of Speech, Language, and Hearing Research, 48*(5), 1065-1079.

Wong, P. (2012). Acoustic characteristics of three-year-olds' correct and incorrect monosyllabic Mandarin lexical tone productions. *Journal of Phonetic.*

Wong, P. (2008). Development of lexical tone production in disyllabic words by 2- to 6-year-old Mandarin-speaking children. Doctoral dissertation, The Graduate Center of the City University of New York, New York.

Xu, C. X., & Xu, Y. (2004). Effects of consonant aspiration on Mandarin tones. *Journal of the International Phonetic Association, 33*(02), 165-181.

Xu, Y. (1997). Contextual tonal variations in Mandarin. *Journal of Phonetics, 25*(1), 61-83.

Xu, Y. (2001a). Fundamental frequency peak delay in Mandarin. *Phonetica*, *58*(1-2), 26-52.

Xu, Y. (2001b). Sources of tonal variations in connected speech. *Journal of Chinese Linguistics, monograph series #17*, 1–31.

Yip, M. (2002). *Tone* (first ed.). United Kingdom: Cambridge University Press.

Zimmerman, I. L., Steiner, V. G., & Pond, R. E. (2002). *Preschool language scale, fourth edition (PLS-4) English edition* (fourth ed.). San Antonio, Texas: Harcourt Assessment, Inc.

Author Note

Puisan Wong, Eye and Ear Institute, Otolaryngology—Head and Neck Surgery, College of Medicine, Ohio State University.

This research was supported by funding from the East Asia and Pacific Summer Institute (EAPSI) of the National Science Foundation (NSF). Part of the study was presented at the 162nd Meeting of the Acoustical Society of America.

Correspondence should be addressed to Puisan Wong at the Eye and Ear Institute, Otolaryngology—Head and Neck Surgery, College of Medicine, The Ohio State University, 915 Olentangy River Road, Columbus, Ohio 43212. Email: pswResearch@gmail.com.

The author thanks Jane Tsay for her assistance in subject recruitment and providing lab space for data collection. Thanks are also extended to 莊淑蘭(Sophia Chuang), 羅美如, and the teachers and staff at Green World Children School and Chung Cheng Nursery School for their generous supports in recruiting participants, the children and parents for their participation in the study, Steve Naber for his input on statistical analyses, and Marisa Monteleone for proof-reading a previous version of the manuscript.

Table 1. Demographic information and language background of the Taiwan children.

ID#	Age	Attending Preschool	Mother Education Level	Father Education Level	Mandarin Language Percentile Score	% of Time Listening to Mandarin	% of Time Listening to Taiwanese	% of Time Listening to English	% of Time Speaking Mandarin	% of Time Speaking Taiwanese	% of Time Speaking English
TC301	2;10	Yes	Master's	Ph.D.	39	90	5	5	100	0	
TC305	3;3	Yes	College	College	59	83	2	15[a]	95	0	5
TC308	3;4	Yes	Junior high school	Senior high school	65	80	20	5	90	10	0
TC310	3;3	Yes	Junior college	Junior college	63	80	20		95	5	
TC312	3;1	Yes	Junior college	Junior college	37	85	5	10[b]	95	0	5
TC313	3;5	Yes	College	College	82	90	10		95	5	
TC315	2;10	Yes	Ph.D.	Ph.D.	64	97	1.5	1.5	99	0.5	0.5
TC318	2;10	No	Junior college	Junior college	39	90	10		95	5	
TC319	2;11	No	College	Ph.D.	30	90	5	5	90	5	5
TC320	3;6	Yes	College	College	78	80	20		95	5	
TC321	3;0	No	College	College	93	90	10		95	5	

[a] With an English-speaking domestic helper at home
[b] From movies and video tapes

Table 2. Background information of the Taiwan and US judges.

Place of Test	ID#	Country of Origin	Age	Gender	Language at Birth	Strongest language	2nd Language	3rd Language	4th Language	Notes
US	UJ01	Taiwan	26;10	F	Mandarin	Mandarin	English	Taiwanese		
US	UJ02	Taiwan	27;6	F	Mandarin	Mandarin	Taiwanese	English		
US	UJ03	Taiwan	29;10	F	Mandarin	Mandarin	English	Taiwanese		
US	UJ04	China	31;3	M	Mandarin	Mandarin	English			
US	UJ06	Taiwan	32;6	M	Mandarin	Mandarin	Taiwanese	English	Japanese, French	
US	UJ07	China	25;0	M	Mandarin	Mandarin	English	Shanghainese	Cantonese	
US	UJ08	China	23;1	M	Shanghainese	Mandarin	Shanghainese	English		Learned Mandarin at 3 years of age
US	UJ09	China	33;8	F	Mandarin	Mandarin	Cantonese	English		
US	UJ11	Taiwan	27;5	F	Mandarin	Mandarin	English			
US	UJ12	China	23;8	F	Mandarin	Mandarin	English			
Taiwan	TJ01	Taiwan	19;3	F	Mandarin	Mandarin	English	Taiwanese	Hakka	
Taiwan	TJ02	Taiwan	19;10	F	Mandarin	Mandarin	Taiwanese	English		
Taiwan	TJ03	Taiwan	19;1	F	Mandarin	Mandarin	Taiwanese	English		
Taiwan	TJ04	Taiwan	18;3	M	Mandarin	Mandarin	Taiwanese	English		
Taiwan	TJ05	Taiwan	23;9	M	Taiwanese	Mandarin	Taiwanese	English		Learned Mandarin at 7 years of age

Table 3. Fleiss' kappa coefficients on the group reliability of the judges

Judges	Tones Produced By	Tone 1	Tone 2	Tone 3	Tone 4	All Tones
15 Taiwan & US Judges	USC3, USA	.80	.66	.52	.78	.69
5 Taiwan judges	USC3, USA	.73	.65	.52	.76	.66
10 US judges	USC3, USA	.84	.68	.52	.78	.71
15 Taiwan & US Judges	USA	.93	.81	.80	.96	.87
5 Taiwan judges	USA	.90	.78	.76	.93	.84
10 US judges	USA	.94	.83	.83	.97	.89
15 Taiwan & US Judges	USC3	.73	.59	.38	.70	.60
5 Taiwan judges	USC3	.64	.59	.40	.69	.58
10 US judges	USC3	.79	.61	.37	.70	.62
5 Taiwan judges	USA, USC3, TWC3	.64	.61	.46	.68	.60
5 Taiwan judges	TWC3	.51	.53	.36	.55	.49

Note: USA, USC3, and TWC3 represent adults in the U.S., three-year-old children in the U.S., and three-year-old children in Taiwan, respectively. The shades of the cells indicate different categories of agreement. The darker the shade indicates the higher the Fleiss' kappa coefficients, and the higher the reliability. Fleiss' kappa of 0-.20, .21-.40, .41-.60, .61-.80, and .81-1 represent slight, fair, moderate, substantial, and almost perfect agreement, respectively.

Table 4. Cohen’s kappa coefficients and percent of agreement of the pairs of U.S. and Taiwan judges on their categorizations of the tones produced by the U.S. adults and children

Judge		TJ01	TJ02	TJ03	TJ04	TJ05	UJ01	UJ02	UJ03	UJ04	UJ06	UJ07	UJ08	UJ09	UJ11	UJ12
	Origin	TW[a]	TW	TW	TW	TW	TW	TW	TW	C[b]	TW	C	C	C	TW	C
		Cohen's kappas on the 290 productions of U.S. adults and children														
TJ01	TW		**.71**	.72	.67	**.60**	.71	.73	.62	.63	.69	.66	.64	.65	.73	.67
TJ02	TW	79		.75	.74	.62	.75	.73	.63	.71	.76	.73	.72	.69	.69	.76
TJ03	TW	79	81		.64	**.58**	.76	.77	.70	.66	.73	.73	.63	.73	.77	.75
TJ04	TW	75	81	73		**.60**	.69	.66	.61	.72	.65	.72	.73	.61	.62	.71
TJ05	TW	**70**	72	**69**	**70**		**.60**	.61	**.54**	.60	**.59**	.65	**.56**	**.60**	**.58**	.64
UJ01	TW	79	81	82	77	**70**		.77	.71	.72	.77	.74	.67	.76	.76	.76
UJ02	TW	80	80	83	74	71	83		.68	.66	.72	.73	.64	.74	.79	.72
UJ03	TW	72	73	78	71	**66**	78	76		.64	.68	.68	**.58**	.68	.69	.69
UJ04	C	72	78	75	79	**71**	79	75	73		.67	.79	.72	.67	.66	.81
UJ06	TW	77	82	80	74	**70**	83	79	76	76		.72	.64	.73	.73	.75
UJ07	C	75	80	80	79	75	81	80	77	84	79		.70	.70	.69	.80
UJ08	C	73	79	72	80	**67**	75	73	**68**	79	73	77		.61	.61	.72
UJ09	C	74	77	80	**71**	**71**	82	81	77	76	80	78	**71**		.73	.70
UJ11	TW	80	77	83	72	**69**	82	84	77	75	80	77	71	80		.69
UJ12	C	75	82	81	79	74	82	79	77	86	81	86	79	78	77	
		Percent Agreement														

Note: . “TJ” represents the five Taiwan judges and “UJ” represents the 10 U.S. judges. The upper half shows the Cohen’s kappa coefficients. The bottom half in shade shows percent of agreement (i.e., percent of productions for which both judges selected the same tone for the productions). The values in bold and highlighted in dark gray mark the bottom 10% of the values. [a]TW represents that the judge’s home country was Taiwan. [b]C represents that the judge’s home country was China. Cohen’s kappa coefficients of 0-.20, .21-.40, .41-.60, .61-.80, and .81-1 represent slight, fair, moderate, substantial, and almost perfect agreement, respectively.

Table 5. Taiwan and U.S. judges' categorization of tones produced by U.S. adults and U.S. and Taiwan children

5a. Tones Produced by U.S. Children

		T1	T2	T3	T4	T1	T2	T3	T4
		Judged by Taiwan Judges (%)				Judged by U.S. Judges (%)			
Target Tone	T1	67	17	4	12	78	11	2	9
	T2	4	67	24	4	5	70	19	6
	T3	3	40	47	11	2	40	44	14
	T4	6	2	13	79	4	2	18	76

5b. Tones Produced by U.S. Adults

		T1	T2	T3	T4	T1	T2	T3	T4
		Judged by Taiwan Judges (%)				Judged by U.S. Judges (%)			
Target Tone	T1	93	6		1	96	4	0	
	T2		96	4		1	96	3	
	T3		18	82			17	83	
	T4	2		4	94			2	98

5c. Tones Produced by Taiwan Children

		T1	T2	T3	T4
		Judged by Taiwan Judges (%)			
Target Tone	T1	71	19	7	3
	T2	10	53	34	3
	T3	9	24	53	15
	T4	15	1	16	68

Table 6. Significant differences in the Taiwan judges' categorization of tone produced by children

Speaker Group	Pattern	χ^2	p-value	Adjusted p-value[a]	df	N	w	Overall Pattern
USC3	T4>T1	8.03	.005	<.05	1	485	.1	
	T4>T2	7.35	.007	<.05	1	480	.1	
	T4>T3	46.24	.000	<.01	1	445	.3	
	T1>T3	20.12	.000	<.01	1	510	.2	
	T2>T3	20.99	.000	<.01	1	505	.2	T4>T1=T2>T3
TWC3	T1>T2	15.94	.000	<.01	1	440	.2	
	T1>T3	14.31	.000	<.05	1	395	.2	
	T4>T2	11.15	.001	<.01	1	455	.2	
	T4>T3	9.97	.002	<.01	1	410	.2	T1=T4>T2=T3

Note: adjusted p-value are the p-values after Bonferroni correction for multiple comparisons.

Table 7. Similarities and Differences in the accuracy of the tones produced by the US and Taiwan children judged by Taiwan judges

	Similarities	Differences	
	Both US and Taiwan Children	US Children	Taiwan Children
Percent Correct of Tones	None of the tones has adult-like accuracy rates Accuracy of T1 and T3 are comparable in both groups		More errors in T2 and T4 than US children
Order of Tone Accuracy	T1&T4 > T3	T4 > T1 = T2 > T3	T4 = T1 > T2 = T3
Error Patterns	T1 mostly →T2 T2 mostly → T3 (more for Taiwan children) T3 mostly → T2, some → T4 T4 mostly → T3	Some T1 → T4	Some T2 → T1 Some T4 → T1