Published in final edited form as:

Int J Audiol. 2017 ; 56(SUP2): S31–S40. doi:10.1080/14992027.2016.1204564.

Validation of list equivalency for Mandarin speech materials to use with cochlear implant listeners

Yongxin Li¹, Shuncheng Wang¹, Qiaodang Su¹, John J. Galvin², and Qian-Jie Fu²

¹Department of Otolaryngology, Head and Neck Surgery, Beijing TongRen Hospital, Capital Medical University, Ministry of Education of China, Beijing, People's Republic of China

²Department of Head and Neck Surgery, David Geffen School of Medicine, UCLA, Los Angeles, CA, USA

Abstract

Objectives—Speech materials validated with normal-hearing listeners may not be appropriate for clinical assessment of cochlear implant (CI) users. The aim of this study was to validate list equivalency of the Mandarin Speech Perception (MSP) sentences, disyllables, and monosyllables in Mandarin-speaking CI patients.

Design—Recognition of MSP sentences, disyllables, and monosyllables each were measured for all 10 lists.

Study sample—67 adult and 32 pediatric Mandarin-speaking CI users.

Results—There was no significant difference between adult and pediatric subject groups for all test materials. Significant differences were observed among lists within each test. After removing one or two lists within each test, no significant differences were observed among the remaining lists. While there was equal variance among lists within a given test, the variance was larger for children than for adults, and increased from monosyllables to disyllables to sentences.

Conclusions—Some adjustment to test lists previously validated with CI simulations was needed to create perceptually equivalent lists for real CI users, suggesting that test materials should be validated in the targeted population. Differences in mean scores and variance across test materials suggest that CI users may differ in their ability to make use of contextual cues available in sentences and disyllables.

Keywords

Cochlear implant; list equivalency; Mandarin speech test

INTRODUCTION

For cochlear implant (CI) users, speech understanding should be assessed using appropriate and validated test materials. Accurate assessment is important for evaluating the efficacy of

an intervention, as well as modifications to a device (e.g., mapping, pre-processing strategies, etc.). Accurate assessment is also important to optimize the device and monitor the progress of auditory rehabilitation. Open-set word and sentence recognition is typically used to assess speech performance of normal-hearing (NH) and hearing-impaired (HI) individuals in clinical and research settings. The tests for the various speech measures are often list-based, but perceptual equivalency among lists is rarely validated in CI users. This can lead to inaccuracies and/or inconsistencies in assessment within and across CI users, depending on which list is used to evaluate performance.

Mandarin Chinese is one of the most popular languages in the world, and there are a growing number of Mandarin-speaking CI users. Different from English, Mandarin Chinese is a tonal language in which syllables with the same vowel-consonant combination can convey different meaning depending on the tone pattern (e.g., Liang, 1963; Lin, 1988). There are four major tonal patterns in Mandarin Chinese, which can be characterized by the variation in the fundamental frequency (F0) contours during voiced speech: Tone 1 (high-flat), Tone 2 (rising), Tone 3 (falling-rising), and Tone 4 (falling). Different from NH listeners, CI users have limited access to F0 cues due to the coarse representation of spectral cues and the lack of fine-structure cues. Voice pitch in CI users can be somewhat conveyed by the spectral pattern (which may not differ substantially across tones) and/or by the temporal envelope used to modulate pulse trains delivered to each electrode (which are poorly perceived in the presence of dynamic spectral envelope cues). Because CI users may not access F0 cues, they depend strongly on other acoustic cues that co-vary with changes in F0 across lexical tones. Vowel duration varies across tones (Fu et al, 1998; Fu & Zeng, 2000), with Tone 3 generally having the longest duration and Tone 4 having the shortest duration. The amplitude contour is also correlated with the F0 contour (Whalen & Xu 1992; Fu & Zeng, 2000). Previous studies with Chinese CI users also reported moderate to good levels of tone recognition, despite limited access to F0 cues (Luo et al, 2008; 2009), suggesting that CI users make good use of the co-varying duration and amplitude contour cues. Tone recognition performance has been shown to correlate with sentence recognition performance in Mandarin-speaking CI listeners (e.g., Fu et al, 2004).

As the population of Chinese CI users continues to grow, there is a great need to develop standardized speech testing materials that are rigorously validated (Ma et al, 2013). Several Mandarin sentence and word test materials have been developed in recent years. The Mandarin Hearing in Noise Test (MHINT) was a first attempt toward developing standardized speech test materials for Mandarin-speaking listeners (Wong et al, 2007). However, there are some limitations for MHINT sentences for use in CI patients. First, the MHINT sentences were not phonetically balanced (in terms of vowels, consonants, or tones) within or across test lists. Second, list equivalency and test-retest variability was validated only in NH listeners and may not hold true for CI users. Third, the MHINT materials are too difficult for pediatric Chinese CI users (Su et al, 2016), who represent the majority of Chinese CI recipients.

Recent studies have shown that Tones 3 and 4 are more easily recognized than Tones 1 or 2 when the F0 contour was removed, as in CI signal processing (e.g., Fu et al, 1998; Luo et al, 2009). Given the importance of lexical tones to Mandarin Chinese speech understanding and

differences in tone perception between NH and CI listeners, it may be more appropriate to validate list equivalency in light of CI signal processing and perception (i.e., with limited spectral and temporal cues). To address these issues, the Mandarin speech perception (MSP) sentence test materials was recently developed and validated by Fu et al. (2011). For the MSP sentences, phonetic balancing was carefully maintained within and across test lists by distributing vowels, consonants, and tones according to the distribution found in everyday Chinese speech (Tang, 1995). List equivalency was validated in NH subjects listening to acoustic simulations of CI processing; no significant difference in performance was observed across MSP sentence lists.

Compared with sentence materials, it is easier to maintain phonetic balancing across Mandarin word lists since there are no contextual constraints. Several Mandarin word test materials have been recently developed, with most using monosyllable words (Han et al, 2009; Ji et al, 2011; Tsai et al, 2009). For example, Han et al (2009) developed Mandarin monosyllabic speech test materials (MSTMs) for speech audiometry. Highly familiar monosyllabic words were phonologically balanced across lists, but list equivalency was validated only in NH listeners. Several research groups have also developed Mandarin disyllable word test materials (Nissen et al, 2005; Wang et al, 2007; Zhu et al, 2012). For example, Wang et al (2007) developed Mandarin disyllabic materials using highly familiar words; materials were phonologically balanced but list equivalence was evaluated using NH listeners. Zhu et al (2012) also developed and evaluated Mandarin disyllable test materials that could be used to evaluate Mandarin-speaking CI users' speech performance. As with the MSP sentences (Fu et al, 2011), vowels, consonants, and tones within and across disyllable word lists were distributed according to their distribution in everyday Chinese speech (Tang, 1995). List equivalency was validated in NH subjects listening to a 4-channel acoustic simulation of CI speech processing.

In the above studies, most of sentence and word materials developed for Mandarin-speaking listeners were evaluated only in NH subjects listening to unprocessed speech, except for Fu et al (2011) and Zhu et al (2012) where list equivalency was validated using acoustic CI simulations. While acoustic simulations are theoretically comparable to CI signal processing, substantial differences in performance have been observed between CI simulations in NH listeners and real CI users. First, variability in performance can be much greater in real CI subjects than in the CI simulations (e.g., Spahr et al., 2012). Second, differences in error patterns can emerge between CI users and CI simulations, especially for tone recognition. For example, Luo et al (2009) compared the tone confusion matrices between CI users and NH subjects listening to 4-channel CI simulation. The poorer CI tone recognition performance was mostly due to confusion between Tones 2 and 3, which resulted in significantly lower performance than with Tones 1 and 4. However, in the CI simulation, the best performance was observed with Tone 3 and the poorest performance with Tone 1. It is possible that CI subjects were unable to make full use of amplitude envelope cues, which have been found to strongly contribute to NH listeners' Tone 3 identification when salient pitch cues are unavailable (Luo & Fu, 2004).

Most speech test lists used for English-speaking CI users also have not been validated for perceptual equivalency using CI simulations and/or real CI users (e.g., HINT sentences from

Nilsson et al, 1994; CUNY sentences from Boothroyd et al, 1985; CNC words from Peterson & Lehiste, 1962). One of the few that has been validated in both CI simulations and real CI users is the AzBio set (Spahr & Dorman, 2004; Sparh et al, 2012), which consists of sentences produced in a conversational speaking style, rather than the clear speaking style as in the HINT or CUNY sentences. Adjustments to equivalent sentence lists in the CI simulation were needed to construct equivalent AzBio sentence lists for CI users. Note that while perceptually equivalent, no effort was made to balance the distribution of phonemes across the AzBio lists, or the number of words per sentence or complexity within and across lists.

Depending on the number of spectral channels available, CI simulation performance may be comparable to that in good CI users, but much better than poor CI users. For Mandarin-speaking CI users, this high variability may impact the relative difficulty across word and sentence test lists. Also, because the errors in tone recognition differ between CI simulations and real CI users (Luo & Fu, 2004), and because tone recognition contributes to sentence recognition performance (e.g., Fu et al, 2004; Chen et al, 2014), list equivalency may be different between NH listeners, CI simulations, and real CI users. Thus, it is important to validate list equivalency in terms of NH performance, but also for real CI users. In this study, list equivalency for MSP sentences, disyllables, and monosyllabic words was evaluated in adult and older pediatric Mandarin-speaking CI users. Adults and children differ in terms of speech pattern development and comprehension, and this may affect performance across tests which may differ in terms of contextual cues (e.g., sentences and disyllables versus monosyllables). Performance for all test lists was compared within and across tests to determine whether performance with a given list was predictive of performance with an alternate list within and across tests.

METHODS

A. Development of Phonetically Balanced Lists

The Mandarin sentence, disyllable and monosyllable materials were developed according to the same criteria described in Fu et al (2011) and Zhu et al (2012): 1) all test materials should be familiar and used in daily life, 2) each test list should be phonetically balanced. To achieve phonetic balance within lists, the targeted distribution of vowels, consonants, and tones was first computed according to the distribution across 3500 commonly used Mandarin Chinese words (Tang, 1995). Due to the limited number of words in each list (70 for sentences, 70 for disyllables, and 50 for monosyllables), some variation of the number of vowels, consonants, and tones was allowed for each list. The number of targeted vowels and consonants within each list was allowed to vary by ± 1 and the number of targeted tones was allowed to vary by ± 2 . No unique word combinations of vowels, consonants, and tones were repeated within a list. The number of words repeated across lists was minimized; disyllables were not repeated across sentence lists and monosyllables were not repeated across disyllable lists. Figure 1 shows the distribution of vowels, consonants, and tones across 3500 commonly used Chinese characters (Tang, 1995) and the present MSP sentence, disyllable, and monosyllable lists.

B. Recordings of Sentence Lists

The same procedure was used to record all of the MSP sentence, disyllable, and monosyllable materials. After developing the phonetically balanced test lists, all materials were clearly produced by a single female talker at a normal speaking rate. At the time of recording, the talker had more than 10 years of professional experience as a broadcaster in a radio station. Each token (sentence, disyllable, or monosyllable) was recorded several times and the most clearly pronounced token was used in the test materials. The mean duration was 1974 ms, 932 ms, and 531 ms for sentence, disyllable, and monosyllable tokens, respectively, with 7 words in each sentence, 2 words in each disyllable, and 1 word in each monosyllable. The mean speaking rate was 3.55, 2.16, and 1.88 words per second (wps) and the mean F0 was 223, 208, and 237 Hz for the sentence, disyllable, and monosyllable tokens, respectively. Audio recordings of all test materials can be downloaded and/or played at the following web site: <http://msp.emilyfufoundation.org/>.

C. Subjects

50 Mandarin-speaking CI users participated in the validation of each of the MSP test materials. For sentences, subjects included 15 pediatric CI users (10 males and 5 females; mean age = 16 years, range = 10 to 19 years) and 35 adult CI patients (22 males and 13 females; mean age = 34 years, range = 21 to 60 years). For disyllables, subjects included 14 pediatric CI users (10 males and 4 females; mean age = 17 years, range = 10 to 19 years) and 36 adult CI patients (21 males and 15 females; mean age = 35 years, range = 21 to 60 years). For monosyllables, subjects included 16 pediatric CI patients (9 males and 7 female; mean age = 17 years, range = 14 to 20 years) and 34 adult CI patients (20 males and 14 females; mean age = 30 years, range = 21 to 59 years). Note that some subjects did not participate in all three tests due to scheduling and availability; as such, a total of 99 CI users participated in the study. Among the pediatric subjects, 12 participated in both the disyllable and sentence tests, but only 1 participated in all three tests. Among the adult subjects, 27 participated in both the disyllable and sentence tests, but only 7 participated in all three tests.

All subjects were recruited from the otolaryngology clinic at Beijing TongRen Hospital in Beijing, China. In this study, individuals under the age of 21 were considered as children in accordance with the National Institutes of Health Policy and Guidelines on the Inclusion of Children as Participants in Research Involving Human Subjects. The minimum age for the present pediatric population was 10 years old to allow for sentence comprehension, computer testing, and a somewhat prolonged test period (testing of all lists required several hours per subject); as such, they can be considered to be “older” pediatric subjects. Table 1 shows demographic information regarding CI experience and duration of deafness for some of the subjects who participated in each of the tests. Unfortunately, not all demographic information was collected for some subjects because those were not regular patients from the otolaryngology clinic and had visited to the clinic from other provinces. As such, there are substantially fewer subjects for some demographic categories than participated in the experiments. Across all tests, the mean CI experience was 5.4 years for children and 1.4 years for adults, the mean duration of deafness was 7.4 years for children and 10.6 years for adults, and the mean CI-aided PTA (across 500, 1000, and 2000 Hz) was 35.3 dB HL for children and 35.6 dB HL for adults. All subjects were paid for their participation, and all

provided informed consent in accordance with the local Institutional Review Board; for pediatric patients, consent was obtained from the parents.

D. Procedures

Stimuli were presented in sound field at 65 dBA via single loudspeaker; subjects were seated directly facing the loudspeaker at a 1 m distance. During testing with each of the speech materials, a list was randomly selected, and sentences or words were randomly selected from within the list (without replacement) and presented to the subject, who repeated what they heard as accurately as possible. The experimenter calculated the percent of words correctly identified. All words in the MSP test materials were scored, resulting in a total of 70 words for each sentence list, 70 words for each disyllable list, and 50 words for each monosyllable list. Each of the test materials contained 10 lists. No training or trial-by-trial feedback was provided during testing. For all materials, all lists were tested with each subject. The test order of the sentence, disyllable, and monosyllable words lists was randomized within and counter-balanced across subjects.

RESULTS

Figure 2 shows scatter plots of word-in-sentence recognition scores for pediatric (filled circles) and adult CI subjects (open circles), as a function of MSP sentence list number. Before statistical analysis, all scores were transformed into rationalized arcsine unit (RAU) scores (Studebaker, 1985) to reduce ceiling and floor performance effects. Table 2 shows the mean RAU score, standard error, and p-value from the Shapiro-Wilk test for normality for each list, for pediatric, adult, and all subjects. For children, non-normal distributions were observed for Lists 2, 3, 7, and 8. For adults, distributions were normal. When all data was combined, non-normal distributions were observed for Lists 2 and 7. A split-plot repeated measures analysis of variance (RM ANOVA) was performed on the data shown in Figure 2, with list as the within-subject factor (1, 2, 3, 4, 5, 6, 7, 8, 9, 10) and subject group as the between-subject factor (pediatric, adult). Mauchly's test showed no violation to sphericity ($p > 0.05$). Results showed a significant effect of list [$F(9,432) = 9.24$, $p < 0.001$], but not for subject group [$F(1,48) = 2.76$, $p = 0.103$]; there was no significant interaction [$F(9,432) = 0.45$, $p = 0.907$]. Table 3 shows the results of Bonferroni pairwise comparisons across lists using pooled data. Results showed that performance was significantly better with List 1 than with Lists 4, 6, 7, 8, and 10, and that performance was significantly better with List 3 than with Lists 4, 7, 8 and 10 ($p < 0.05$ in all cases). When Lists 1 and 3 were excluded (the lists with the top two mean scores), Bonferroni pairwise comparisons showed no significant differences among lists ($p > 0.05$ in all cases).

Figure 3 shows scatter plots of disyllable recognition scores for pediatric (filled circles) and adult CI subjects (open circles), as a function of MSP disyllable list number. Before statistical analysis, all scores were transformed into RAU units to reduce ceiling and floor performance effects. Table 4 shows the mean RAU score, standard error, and p-value from the Shapiro-Wilk test for normality for each list, for pediatric, adult, and all subjects. A non-normal distribution was observed only for adult subjects with List 7. A split-plot RM ANOVA was performed on the data shown in Figure 3, with list as the within-subject factor

and subject group as the between-subject factor. Mauchly's test showed no violation to sphericity ($p > 0.05$). Results showed a significant effect of list [$F(9,432) = 4.12, p < 0.001$], but not for subject group [$F(1,48) = 0.16, p = 0.689$]; there was no significant interaction [$F(9,432) = 1.04, p = 0.402$]. Table 5 shows the results of Bonferroni pairwise comparisons across lists using pooled data. Results showed that performance was significantly better with List 2 than with Lists 1, 3, and 10 ($p < 0.05$ in all cases). When Lists 2 and 3 were excluded (the lists with the highest and lowest mean scores, respectively), Bonferroni pairwise comparisons showed no significant differences among the remaining lists ($p > 0.05$ in all cases).

Figure 4 shows scatter plots of monosyllable recognition scores for pediatric (filled circles) and adult CI subjects (open circles), as a function of MSP monosyllable list number. Before statistical analysis, all scores were transformed into RAU units to reduce ceiling and floor performance effects. Table 6 shows the mean RAU score, standard error, and p-value from the Shapiro-Wilk test for normality for each list, for pediatric, adult, and all subjects. Non-normal distributions were observed in children for List 1 and in adults for Lists 2 and 7. A split-plot RM ANOVA was performed on the data shown in Figure 4, with list as the within-subject factor and group as the between-subject factor. Mauchly's test showed no violation to sphericity ($p > 0.05$). Results showed a significant effect of list [$F(9,432) = 4.09, p < 0.001$], but not for subject group [$F(1,48) = 0.04, p = 0.852$]; there was a significant interaction [$F(9,432) = 3.87, p < 0.001$]. Table 7 shows the results of Bonferroni pairwise comparisons across lists using pooled data. Results showed that performance was significantly better with List 1 than with Lists 3, 5, 6, and 10 ($p < 0.05$ in all cases). When List 1 was excluded (the list with the highest mean score), Bonferroni pairwise comparisons showed no significant differences among the remaining lists ($p > 0.05$ in all cases).

To test list reliability, correlation analyses were performed among all lists within each of the sentence, disyllable, and monosyllable tests. Separate correlations were performed for the pediatric and adult subjects using raw scores. Bonferroni adjustments to the p-value were applied to correct for family-wise error. For children, there was a significant correlation among all sentence lists, all disyllable lists, and all monosyllable lists ($r > 0.85, p < 0.001$ in all cases). For adults, there was a significant correlation among all sentence lists, all disyllable lists, and all monosyllable lists ($r > 0.90, p < 0.001$ in all cases).

To examine the relationship of lists among the different tests, correlational analyses were performed between each sentence and each disyllable list, between each sentence and monosyllable list, and between each disyllable and monosyllable list. Separate correlations were performed for the pediatric and adult subjects using raw scores from only the subjects who completed both tests within each comparison. Accordingly, data were compared for 12 pediatric subjects between sentence and disyllable lists; no comparison was made between monosyllable and disyllable or sentence test lists because only 1 subject completed all three tests. For adult subjects, data were compared between sentence and disyllable tests for the 29 subjects who completed both tests, and between monosyllable and disyllable or sentence tests for the 7 subjects who completed all three tests. Bonferroni adjustments to the p-value were applied to correct for family-wise error. For pediatric subjects ($n=12$), significant correlations were observed between all sentence and disyllable sentence lists ($p < 0.005$),

except for sentence List 2 versus disyllable List 10 ($p = 0.0056$); r^2 values for cross-test list comparisons ranged from 0.55 (sentence List 2 versus disyllable List 10) to 0.86 (sentence List 6 versus disyllable List 9). For adult subjects ($n=29$), significant correlations were observed between all sentence and disyllable sentence lists ($p < 0.005$); r^2 values ranged from 0.80 (sentence List 4 versus disyllable List 4) to 0.95 (sentence List 2 versus disyllable List 8). However, there was no significant correlation for adults ($n=7$) between any of the monosyllable lists and any of the sentence or disyllable lists ($p > 0.05$ in all cases).

DISCUSSION

The present study showed that some adjustment was needed to maintain perceptual equivalency across MSP sentence, disyllable, and monosyllable lists for real CI users, even though list equivalency had been previously established for unprocessed speech or CI simulations in NH listeners. However, it is possible to achieve list equivalency for CI users by selecting subsets of lists. By doing so, the MSP sentence, disyllable, and monosyllable materials meet four important criteria for development of speech test materials for CI users: familiarity, homogeneity, phonetic balance, and list equivalency (Tsai et al, 2009). The present results suggest that list equivalency for newly developed or existing materials should be validated in the targeted population. This is especially true for CI users, who tend to exhibit more variability in performance within and across subjects than NH listeners.

After converting to RAU scores, significant differences were observed among sentence, disyllable, and monosyllable lists. Equivalent lists for each speech test were established using Bonferroni-corrected pairwise comparisons of pooled pediatric and adult data. With this approach, equivalent lists were created for the MSP sentences (Lists 2, 4, 5, 6, 7, 8, 9, and 10), disyllables (Lists 1, 4, 5, 6, 7, 8, 9, and 10), and monosyllables (Lists 2–10). One source of variability across lists may have been the distribution of vowels, consonants, and tones across lists. While phonetic balancing was carefully considered for each MSP sentence list, some variation in the distribution may have occurred across lists due to contextual constraints associated with sentence materials.

Another approach to create list equivalence would have been to combine lists and/or sentences, as done with the English Az-Bio sentences (Spahr et al, 2004; 2012). However, this would have resulted in a maximum of only 5 lists for the present MSP sentence materials, arguably too few to be used in a clinical or research setting where multiple lists must be tested. It would also be difficult to generate new lists while maintaining phonetic balance across lists, as well as familiarity, homogeneity, as recommended by Tsai et al (2009). The disyllable and monosyllable lists each contain 35 words and 50 words, respectively, so combining lists would not be an option if reasonable test duration is considered. As such, removing the tests lists that produced significantly different performance was the best option to obtain a sufficient number of MSP sentence, disyllable, and monosyllable lists with which to test CI users.

Yet another approach would have been to normalize lists according to a psychometric function as in Nissen et al (2005). In that study, NH disyllable performance was measured in quiet at various presentation levels, and level adjustments were made to disyllables to create

perceptual equivalency across lists. One difference between the MSP materials and other Mandarin test materials is that the MSP materials were phonetically balanced across lists. Without phonetic balancing, it would be possible to create lists that were perceptually equivalent, but may not necessarily reflect the distribution of tones and syllables in Mandarin Chinese. Also, given the effects of the AGC, microphone sensitivity, and amplitude mapping in CI users that may interact with stimulus level presentation, testing in quiet at different presentation levels may not produce consistent performance across real CI subjects. Similar level-adjustment approaches have been used for the MHINT (Wong et al, 2007) and HINT sentences (Nilsson et al, 1994) in noise with NH listeners. However, as the present data indicate, testing in noise may not have been possible for many of the present real CI subjects. The level adjustments derived from NH performance in noise may not result in list equivalency for testing CI users in quiet.

For the MSP sentences, mean recognition across all lists and subjects was lower and more variable than observed with the 4-channel CI simulations in Fu et al (2011), in which only adult NH subjects were tested. Mean performance (across lists) for the top one-third of adult CI users in this study was 91.0% correct (range = 83.3% to 98.0% correct), comparable to that in Fu et al (2011). Mean performance for the bottom one-third was 24.8% correct (range = 6.4% to 40.2% correct), indicating the variability in performance that may occur when testing real CI users. For adult subjects, the distribution of RAU scores was normal. In pediatric subjects, non-normal distributions were observed for 4 of the 10 lists. It is possible that a greater number of pediatric subjects might have resulted in normal distributions for all lists, as observed with the adult subjects.

List equivalency for the MSP disyllables was originally validated with adult NH subjects listening to 4-channel CI simulations (Zhu et al, 2012). In this study, mean recognition across all lists and CI subjects was much lower and more variable (mean = 60.9% correct; range: 1.4% to 98.6% correct) than observed with the 4-channel CI simulations (mean = 78.3% correct; range 76.4% to 79.3% correct). This discrepancy reflects differences between real CI users and NH subjects listening to CI simulations, who tend to be more homogenous in response. Also, CI simulations generally characterize only signal degradation and do not model factors that may significantly affect real CI performance (e.g., duration of deafness, age at implantation, CI experience, etc.). Note that mean recognition (across lists) for the top-third of the present adult CI subjects was 86.0% correct (range = 78.3% to 98.4% correct), somewhat higher than observed with the NH adults in the CI simulation. In Zhu et al (2012), mean disyllable recognition was significantly poorer than mean sentence recognition with the CI simulation for a subset of subjects who completed both tests. In this study, mean disyllable recognition was slightly better than mean sentence recognition for CI subjects who completed both tests.

Unlike the MSP sentences (Fu et al, 2011) and disyllables (Zhu et al, 2012), monosyllable list equivalency was not validated using CI simulations. Mean monosyllable recognition (across all CI subjects and test lists) was significantly lower than that for sentences or disyllables, even when only the top one-third of adult CI subjects is considered. It is interesting to note that while mean performance increased with stimulus complexity (from monosyllables to disyllables to sentences), so did the variance. This suggests that while CI

users may be able to effectively use context cues available in sentences and disyllables in general (i.e., a more central pattern recognition process), there is greater variability among CI users' ability to use these cues. In contrast, there was lower recognition performance and less variance in monosyllable scores, which may reflect difficulties more related to the degraded speech patterns with the CI (i.e., a more peripheral process).

Within a given test, all lists were significantly correlated, suggesting good alternate forms reliability. Correlation analyses were also performed across tests among all lists for subjects who completed both tests. For pediatric ($n=12$) and adult CI subjects ($n=27$), all sentence lists were significantly correlated with all disyllable lists, suggesting that any of the sentence lists were highly predictive of any of the disyllable lists. This is not surprising, as all lists within and across tests were balanced in terms of the distribution of vowels, consonants, and tones. For adult CI subjects ($n=7$), there was no significant correlations between any of the monosyllable lists and any of the sentence or disyllable lists. Unfortunately, only a few adult subjects completed all three tests; it is unclear whether a greater number of subjects would have produced different results. However, even within this small subset of subjects, all sentence and disyllables lists were significantly correlated ($p < 0.005$), suggesting that performance with monosyllables might not be predictive of performance with disyllables or sentences. Overall, the results suggest that monosyllable testing should be performed in all subjects, but that either sentence or disyllable testing would be sufficient as an additional speech test.

Mean adult and pediatric performance with the MSP sentences was also much poorer than observed in Su et al (2016) with MSP sentences. Note that the recording was different between this and the Su et al (2016) study; also, 10 lists were tested in the present study but only 1 list in Su et al (2016). The number of adult and pediatric CI subjects also differed between studies, with 34–36 adults and 14–16 children for each test in this study versus 15 adults and 11 children in Su et al (2016). Most notably, the range of performance was much greater for the present adult and pediatric CI subjects than in Su et al (2016). In this study, mean sentence recognition (across all lists) ranged from 1.4% to 100% correct for adults, and from 1.4% to 100% correct for children. In Su et al (2016), sentence recognition scores ranged from 44.3% to 100% correct for adults, and from 40.0% to 100% correct for children. Clearly, the range of performance was much larger in this study than in Su et al (2016), perhaps because of differing amounts of CI experience and duration of deafness among subjects in each study. When the top-third of the present adult and CI subjects are considered, sentence recognition performance was more comparable across studies (adult mean = 86.0% correct, range = 78.3% to 95.4% correct; pediatric mean = 80.8% correct, range = 54.9% to 92.6% correct).

In this study, sentence, disyllable and monosyllable recognition was measured in quiet and list equivalency was established using data collected in quiet. It is unclear whether list equivalency would hold in noise. The present distribution of performance in quiet suggests that it may be difficult to measure performance in noise for many CI subjects. The distribution of performance may also be affected in noise, with some subjects able to tolerate some amount of noise and others unable to recognize sentences even in quiet. The best approach would be to re-evaluate list equivalency in noise; it is possible that different lists

may be appropriate for testing CI performance in quiet or in noise. In this study, performance with each list was analyzed using RAU scores, and list equivalency was established using post-hoc pairwise comparisons with Bonferroni corrections for family-wise error. Minimal differences in performance across lists were likely due to the careful balancing of vowels, consonants and lexical tones across lists, as well as using familiar words for sentence, disyllable, and monosyllable testing. Nonetheless, it would be valuable to further validate list equivalency using different subjects at multiple research and clinical sites. The pediatric subjects in this study were older (between 10 and 20 years old). It is unclear whether performance may differ among test lists or whether the present materials are age-appropriate for a younger pediatric CI population. Finally, the perceptually equivalent lists reported in this study should be validated with a different set of CI subjects from multiple clinics for further validation.

CONCLUSIONS

The present study reports several findings regarding issues of list equivalency for newly developed or existing test materials used to evaluate Mandarin-speaking CI users:

1. Validation of speech test materials with NH subjects, even when listening to a CI simulation, is not sufficient to establish list equivalency for CI users. Along with word familiarity and phonemic balancing, it is necessary to validate test materials using the target populations. In the present study, statistical analyses were used to identify perceptually equivalent lists (in terms of RAU scores) that could be used to test adult and older pediatric Chinese CI users.
2. There was a strong correlation between all disyllable and sentence test lists, suggesting that either measure may be sufficient to assess CI performance in quiet. However, there was no correlation between monosyllable and sentence or disyllable recognition, suggesting that monosyllable recognition should also be measured in all CI subjects.
3. After transforming to RAU scores, there was no significant difference in mean performance (across lists) between the present pediatric and adult CI subjects for any of the tests. While mean scores increased from monosyllables to disyllables to sentences, so did the variance, suggesting differences among CI subjects to make use of contextual cues available with disyllables and sentences.

Acknowledgments

The authors thank all the subjects who participated in this study. This work was partially supported by NIH grant DC004993.

References

- Bench J, Kowal A, Bamford J. The BKB (Bench-Kowal-Bamford) sentence lists for partially-hearing children. *Br J Aud.* 1979; 13:108–112.
- Boothroyd, A., Hanin, L., Hnath, T. A sentence test of speech perception: Reliability, set equivalence, and short term learning (internal report RCI 10). New York: City University of New York; 1985.

- Chen Y, Wong LL, Chen F, Xi X. Tone and sentence perception in young Mandarin-speaking children with cochlear implants. *Int J Pediatr Otorhinolaryngol*. 2014; 78:1923–1930. [PubMed: 25213422]
- Fu Q-J, Zeng F-G. Effects of envelope cues on Mandarin Chinese tone recognition. *Asia-Pacific J Speech Lang Hear*. 2000; 5:45–57.
- Fu Q-J, Zeng F-G, Shannon RV, Soli SD. Importance of tonal envelope cues in Chinese speech recognition. *J Acoust Soc Am*. 1998; 104:505–510. [PubMed: 9670541]
- Fu Q-J, Hsu C-J, Horng M-J. Effects of Speech Processing Strategy on Chinese Tone Recognition by Nucleus-24 Cochlear Implant Users. *Ear Hear*. 2004; 25:501–508. [PubMed: 15599196]
- Fu QJ, Zhu M, Wang X. Development and validation of the Mandarin speech perception test. *J Acoust Soc Am*. 2011; 129:EL267–EL273. [PubMed: 21682363]
- Han D, Wang S, Zhang H, Chen J, Jiang W, Mannell R, et al. Development of Mandarin monosyllabic speech test materials in China. *Int J Audiol*. 2009; 48:300–311. [PubMed: 19842805]
- Ji F, Xi X, Chen AT, Zhao WL, Zhang X, Ni YF, et al. Development of a Mandarin monosyllable test material with homogenous items (II): lists equivalence evaluation. *Acta Otolaryngol*. 2011; 131:1051–1060. [PubMed: 21599549]
- Li Y, Zhang G, Kang H-Y, Liu S, Han D, Fu Q-J. Effects of speaking style on speech intelligibility for Mandarin-speaking cochlear implant users. *J Acoust Soc Am*. 2011; 129:EL242–EL247. [PubMed: 21682359]
- Liang ZA. The auditory perception of Mandarin tones. *Acta Physica Sinica*. 1963; 26:85–91.
- Lin M-C. The acoustic characteristics and perceptual cues of tones in Standard Chinese. *Chinese Yuwen*. 1988; 204:182–193.
- Luo X, Fu Q-J. Enhancing Chinese tone recognition by manipulating amplitude envelope: implications for cochlear implants. *J Acoust Soc Am*. 2004; 116(6):3659–3667. [PubMed: 15658716]
- Luo X, Fu Q-J, Galvin JJ. Vocal emotion recognition by normal-hearing listeners and cochlear implant users. *Trends Amplif*. 2008; 11:301–315.
- Luo X, Fu Q-J, Wu HP, Hsu C-J. Concurrent-vowel and tone recognition by Mandarin-speaking cochlear implant users. *Hear Res*. 2009; 256:75–84. [PubMed: 19595753]
- Ma X, McPherson B, Ma L. Chinese speech audiometry material: Past, present, future. *Hear Balance Communication*. 2013; 11:52–56.
- Nilsson M, Soli SD, Sullivan JA. Development of the Hearing in Noise Test for the measurement of speech reception thresholds in quiet and in noise. *J Acoust Soc Am*. 1994; 95:1085–1099. [PubMed: 8132902]
- Nissen SL, Harris RW, Jennings LJ, Eggett DL, Buck H. Psychometrically equivalent Mandarin bisyllabic speech discrimination materials spoken by male and female talkers. *Int J Audiol*. 2005; 44:379–390. [PubMed: 16136788]
- Peterson GE, Lehiste I. Revised CNC lists for auditory tests. *J Speech Hear Disorders*. 1962; 27:62–70.
- Spahr AJ, Dorman MF. Performance of subjects fit with the Advanced Bionics CII and Nucleus 3G cochlear implant devices. *Arch Otolaryngol Head Neck Surg*. 2004; 130:624–628. [PubMed: 15148187]
- Spahr AJ, Dorman MF, Litvak LM, Van Wie S, Gifford RH, Loizou PC, et al. Development and validation of the AzBio sentence lists. *Ear Hear*. 2012; 33:112–117. [PubMed: 21829134]
- Studebaker GA. A "rationalized" arcsine transform. *J Speech Hear Res*. 1985; 28:455–462. [PubMed: 4046587]
- Su Q, Galvin JJ, Zhang G, Li Y, Fu QJ. Influence of speech variations on speech intelligibility by mandarin-speaking adult and pediatric cochlear implant patients. *Trends Hear*. 2016 (in review).
- Tang YH. Statistical analysis of Mandarin Chinese. *J Chengde Teachers' College Nationalities*. 1995:66–76.
- Tsai KS, Tseng LH, Wu CJ, Young ST. Development of a mandarin monosyllable recognition test. *Ear Hear*. 2009; 30:90–99. [PubMed: 19125031]
- Wang S, Mannell R, Newall P, Zhang H, Han D. Development and evaluation of Mandarin disyllabic materials for speech audiometry in China. *Int J Audiol*. 2007; 46:719–731. [PubMed: 18049961]

- Whalen DH, Xu Y. Information for Mandarin tones in the amplitude contour and in brief segments. *Phonetica*. 1992; 49:25–47. [PubMed: 1603839]
- Wong LL, Soli SD, Liu S, Han N, Huang MW. Development of the Mandarin Hearing in Noise Test (MHINT). *Ear Hear*. 2007; 28(2 Suppl):70S–74S. [PubMed: 17496652]
- Yin, B., Felley, M. Chinese Romanization: Pronunciation and Orthography. Beijing: Sinolingua; 1990.
- Zhu M, Fu QJ, Galvin JJ 3rd, Jiang Y, Xu J, Xu C, et al. Mandarin Chinese speech recognition by pediatric cochlear implant users. *Int J Pediatr Otorhinolaryngol*. 2011; 75:793–800. [PubMed: 21489643]
- Zhu M, Wang X, Fu QJ. Development and validation of the Mandarin disyllable recognition test. *Acta Otolaryngol*. 2012; 132:855–861. [PubMed: 22768800]

Figure 1.

Distribution (in percent) of 35 vowels (Panel A), 21 consonants (Panel B), and 5 tones (Panel C) across 3500 commonly used Chinese words (from Tang, 1995) and for MSP sentences, disyllables, and monosyllables. Note that “None” in Panel B indicates the distribution of Chinese words that have no initial consonants. Tones 0, 1, 2, 3, and 4 represent neutral, flat, rising, falling-rising, and falling tones, respectively. Error bars show the standard error of the percent distribution across lists.

Figure 2. Scatterplots of sentence recognition scores for pediatric (filled circles) and adult CI subjects (open circles) for the different test lists. The solid lines indicate mean scores across all subjects.

Figure 3. Scatterplots of disyllable recognition scores for pediatric (filled circles) and adult CI subjects (open circles) for the different test lists. The solid lines indicate mean scores across all subjects.

Figure 4.

Scatterplots of monosyllable recognition scores for pediatric (filled circles) and adult CI subjects (open circles) for the different test lists. The solid lines indicate mean scores across all subjects.

Demographic information (where available) for pediatric and adult CI subjects. The total n shows the total number of subjects who participated in each test. For the mean and range of CI experience, duration of deafness, and pure tone average (PTA across 500, 1000, and 2000 Hz), the n shows the number of subjects for which demographic information was available.

Table 1

	total n	CI experience (yrs)		Duration of deafness (yrs)		PTA (dB HL)	
		Mean	range	Mean	range	Mean	range
Children	Sentence	5.1 (n=7)	1– 11.7	6.4 (n=6)	3.3– 10	34.9 (n=15)	23.3– 56.7
	Disyllable	4.9 (n=7)	1– 11.7	8.2 (n=6)	3.3– 15.7	35.5 (n=14)	25– 56.7
	Monosyllable	6.2 (n=15)	0.4– 13.2	7.5 (n=13)	1– 15	35.5 (n=16)	25– 35.5
Adults	Sentence	1.4 (n=29)	0.1– 5	9.6 (n=27)	1– 30	35.6 (n=35)	18.3– 50
	Disyllable	1.6 (n=29)	0.1– 5	11 (n=28)	0.2– 30	37.1 (n=36)	23.3– 50
	Monosyllable	2.5 (n=33)	0.2– 15.5	11.1 (n=29)	1– 33	33 (n=28)	15– 51.7

Table 2

Mean RAU scores, standard error (SE), and p-value from Shapiro-Wilk test for normality of distribution (Norm.) for each sentence list, for pediatric, adult, and all subjects.

List	Child			Adult			All		
	Mean	SE	Norm.	Mean	SE	Norm.	Mean	SE	Norm.
1	56.55	8.43	0.081	71.01	5.85	0.502	66.67	4.86	0.085
2	51.21	6.97	0.007*	66.27	5.37	0.676	61.75	4.38	0.010*
3	52.76	8.36	0.045*	71.39	6.12	0.294	65.80	5.07	0.039
4	46.10	8.34	0.501	62.07	5.84	0.734	57.28	4.86	0.489
5	48.00	8.27	0.312	66.05	5.66	0.964	60.64	4.78	0.424
6	48.90	8.22	0.137	63.78	5.47	0.767	59.31	4.61	0.112
7	45.22	7.97	0.013*	62.63	5.76	0.468	57.41	4.78	0.017*
8	45.11	7.76	0.128	61.83	5.53	0.590	56.82	4.60	0.208
9	51.38	9.13	0.021*	67.05	5.61	0.723	62.35	4.85	0.061
10	43.64	8.14	0.292	62.73	5.93	0.899	57.00	4.93	0.255

Non-normal distributions (p < 0.05) are indicated by asterisks.

Table 3

Results of post-hoc Bonferroni pairwise comparisons among sentence lists.

List	2	3	4	5	6	7	8	9	10
1	0.12	1.00	<0.01*	0.02	0.01*	<0.01*	<0.01*	0.30	<0.01*
2		0.90	0.25	1.00	1.00	0.24	0.26	1.00	0.08
3			<0.01*	0.06	0.12	<0.01*	<0.01*	1.00	<0.01*
4				1.00	1.00	1.00	1.00	0.43	1.00
5					1.00	1.00	1.00	1.00	1.00
6						1.00	1.00	1.00	1.00
7							1.00	0.21	1.00
8								0.13	1.00
9									0.19

Significant differences ($p < 0.05$) are indicated by asterisks.

Table 4

Mean RAU scores, standard error (SE), and p-value from Shapiro-Wilk test for normality of distribution (Norm.) for each disyllable list, for pediatric, adult, and all subjects.

List	Child			Adult			All		
	Mean	SE	Norm.	Mean	SE	Norm.	Mean	SE	Norm.
1	57.24	6.30	0.386	59.52	4.16	0.262	58.88	3.44	0.193
2	60.95	6.87	0.562	65.66	4.07	0.105	64.34	3.48	0.053
3	52.52	6.04	0.943	61.63	4.00	0.076	59.08	3.36	0.082
4	58.35	6.37	0.903	63.88	4.65	0.050	62.34	3.77	0.058
5	58.51	7.14	0.491	64.99	4.40	0.274	63.17	3.73	0.164
6	56.47	5.15	0.910	64.49	4.39	0.109	62.24	3.48	0.200
7	56.44	5.70	0.987	65.91	4.58	0.035 *	63.26	3.68	0.082
8	56.12	7.14	0.924	63.52	4.15	0.295	61.45	3.59	0.313
9	57.44	7.45	0.659	64.83	4.70	.0459	62.76	3.96	0.372
10	52.16	6.53	0.666	60.83	4.64	0.083	58.40	3.82	0.081

Non-normal distributions (p < 0.05) are indicated by asterisks.

Table 5

Results of post-hoc Bonferroni pairwise comparisons among disyllable lists.

List	1	2	3	4	5	6	7	8	9	10
1		<0.00*	1.00	1.00	0.59	1.00	1.00	1.00	1.00	1.00
2			<0.00*	1.00	1.00	1.00	1.00	0.79	1.00	<0.00*
3				0.32	0.06	1.00	0.23	1.00	0.62	1.00
4					1.00	1.00	1.00	1.00	1.00	0.85
5						1.00	1.00	1.00	1.00	0.11
6							1.00	1.00	1.00	0.50
7								1.00	1.00	0.31
8									1.00	0.39
9										0.09

Significant differences ($p < 0.05$) are indicated by asterisks.

Mean RAU scores, standard error (SE), and p-value from Shapiro-Wilk test for normality of distribution (Norm.) for each monosyllable list, for pediatric, adult, and all subjects.

Table 6

List	Child			Adult			All		
	Mean	SE	Norm.	Mean	SE	Norm.	Mean	SE	Norm.
1	58.12	5.20	0.040 *	54.96	3.86	0.290	55.91	3.10	0.193
2	56.01	5.79	0.253	51.83	4.02	0.046 *	53.08	3.29	0.053
3	47.46	6.27	0.130	53.63	3.82	0.094	51.78	3.26	0.082
4	54.40	6.30	0.177	53.60	3.98	0.059	53.84	3.33	0.058
5	51.12	6.53	0.098	53.00	3.86	0.356	52.44	3.30	0.164
6	49.11	5.76	0.722	52.98	3.89	0.170	51.82	3.21	0.200
7	54.56	5.88	0.256	52.78	3.49	0.041 *	53.31	2.98	0.082
8	51.80	6.21	0.312	56.09	3.95	0.183	54.80	3.31	0.313
9	50.15	5.02	0.093	53.69	3.68	0.521	52.63	2.96	0.372
10	50.16	5.89	0.072	53.15	4.08	0.092	52.25	3.33	0.081

Non-normal distributions (p < 0.05) are indicated by asterisks.

Table 7

Results of post-hoc Bonferroni pairwise comparisons among monosyllable lists.

List	1	2	3	4	5	6	7	8	9	10
1		1.00	<0.01*	1.00	0.01*	0.00*	0.91	1.00	0.06	0.01*
2			0.58	1.00	1.00	0.62	1.00	1.00	1.00	1.00
3				0.16	1.00	1.00	0.68	0.82	1.00	1.00
4					1.00	1.00	1.00	1.00	1.00	1.00
5						1.00	1.00	1.00	1.00	1.00
6							1.00	1.00	1.00	1.00
7								1.00	1.00	1.00
8									1.00	1.00
9										1.00

Significant differences ($p < 0.05$) are indicated by asterisks.