


# Development and Validation of a Mandarin Chinese Adaptation of AzBio Sentence Test (CMnBio)

Trends in Hearing  
Volume 26: 1–12  
© The Author(s) 2022  
Article reuse guidelines:  
sagepub.com/journals-permissions  
DOI: 10.1177/23312165221134007  
journals.sagepub.com/home/tia  


Xin Xi<sup>1,2,\*</sup> , Ye Wang<sup>3,\*</sup>, Ya Shi<sup>4</sup>, Rui Gao<sup>5</sup>, Siqi Li<sup>6</sup> , Xinyue Qiu<sup>4</sup>, Qian Wang<sup>1,2</sup> and Li Xu<sup>7</sup> 

## Abstract

A new sentence recognition test in Mandarin Chinese was developed and validated following the principles and procedures of development of the English AzBio sentence materials. The study was conducted in two stages. In the first stage, 1,020 sentences spoken by 4 talkers (2 males and 2 females) were processed through a 5-channel noise vocoder and presented to 17 normal-hearing Mandarin-speaking adults for recognition. A total of 600 sentences (150 from each talker) in the range of approximately 62 to 92% correct (mean = 78.0% correct) were subsequently selected to compile 30, 20-sentence lists. In the second stage, 30 adult CI users were recruited to verify the list equivalency. A repeated-measures analysis of variance followed by the post hoc Tukey's test revealed that 26 of the 30 lists were equivalent. Finally, a binomial distribution model was adopted to account for the inherent variability in the lists. It was found that the inter-list variability could be best accounted for with a 65-item binomial distribution model. The lower and upper limits of the 95% critical differences for one- and two-list recognition scores were then generated to provide guidance for detection of a significant difference in recognition scores in clinical settings. The final set of 26 equivalent lists contains sentence materials more difficult than those found in other speech audiometry materials in Mandarin Chinese. This test should help minimize the ceiling effects when testing sentence recognition in Mandarin-speaking CI users.

## Keywords

speech audiometry, sentence recognition, cochlear implantation, Mandarin Chinese, adults

Received 16 July 2022; Revised received 2 October 2022; accepted 3 October 2022

## Introduction

Speech audiometry refers to testing of one's auditory function using speech stimuli. It is widely used for hearing assessment and rehabilitative outcome measurement. As speech stimuli are most comparable to daily communication, speech recognition performance is regarded as a critical measure in providing evidence-based best practice for cochlear implant (CI) recipients (Messersmith et al., 2019). In CI clinical practice, speech perception testing is required for the determination of candidacy, assessment of performance outcomes, and evaluation of intervention efficacy. Commonly used speech materials include both word [e.g., the Consonant-Nucleus-Consonant (CNC) test (Peterson & Lehiste, 1962), the Central Institute for the Deaf (CID) Auditory Test W-22 (Hirsh et al., 1952), and the Northwestern University Auditory Test No. 6 (NU-6) (Tillman & Carhart, 1966)] and sentence recognition

<sup>1</sup>Department of Otolaryngology, Head & Neck Surgery, The Sixth Medical Center, Chinese PLA General Hospital, Beijing, China

<sup>2</sup>National Clinical Research Center for Otolaryngologic Diseases, Beijing, China

<sup>3</sup>Department of Otolaryngology, Zhejiang Hospital, Hangzhou, Zhejiang, China

<sup>4</sup>School of Medical Technology, Zhejiang Chinese Medical University, Hangzhou, Zhejiang, China

<sup>5</sup>School of BioMedical Engineering, Capital Medical University, Beijing, China

<sup>6</sup>School of Communication Science, Beijing Language and Culture University, Beijing, China

<sup>7</sup>Communication Sciences and Disorders, Ohio University, Athens, OH, USA

\*These authors are co-first authors who have contributed equally to this work.

### Corresponding Author:

Li Xu, Communication Sciences and Disorders, Ohio University, Athens, OH 45701, USA.  
Email: xul@ohio.edu



materials [e.g., the Hearing in Noise Test (HINT) sentences (Nilsson et al., 1994), the City University of New York (CUNY) sentences (Boothroyd et al., 1985), and the AzBio sentences (Spahr et al., 2012)].

As CI technology has improved and individuals with more residual hearing have been identified as CI candidates, CI recipients have demonstrated remarkable improvement in speech perception. Consequently, ceiling effects have been observed on certain sentence materials when administered in quiet (Bassim et al., 2005; Gifford et al., 2008). Gifford et al. (2008) found that as many as 71% of the adult CI subjects scored 85% correct or higher for the HINT sentences in quiet, and that 28% of the subjects scored maximum performance of 100% correct. Bassim et al. (2005) reported that the average sentence recognition scores of postlingually-deafened CI recipients reached 96% and 87% correct with the CUNY sentences and the HINT sentences, respectively, one year after implantation.

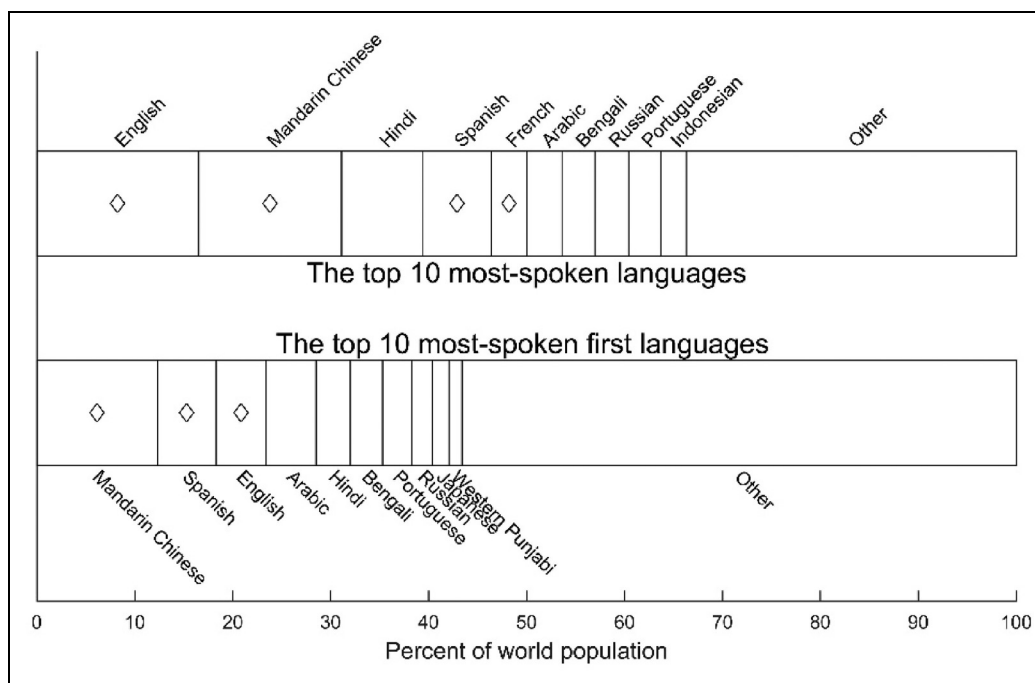
The need for more difficult sentence materials for clinical testing was met with the development of the AzBio sentences (Spahr et al., 2012). To achieve a higher level of difficulty, the English AzBio sentences were recorded by multiple talkers (two males and two females) using a conversational speaking style with longer sentence lengths and limited contextual cues. The English AzBio sentences contained 33 lists of 20 sentences, and 29 of the 33 lists had been verified to be of equivalent intelligibility based on the recognition scores of 15 adult CI listeners. In a separate study, Gifford et al. (2008) found that only one out of the 156 postlingually-deafened CI recipients scored 100% correct with the English AzBio sentences. Given the advantage of avoiding the ceiling effects in quiet over other sentence materials, the AzBio sentence test was included in the 2011 Minimum Speech Test Battery that serves as the recommendation for evaluation of pre- and post-implantation speech recognition in adults in the United States (MSTB, 2011). Fifteen of the 29 lists that produced the most similar performance were selected to form the commercially available AzBio Sentence Test, and the next eight lists with the most similar performance scores were included in the Minimum Speech Test Battery for adult CI users (MSTB, 2011). Since the advent of the English AzBio sentences, similar sentences were developed and validated in several other languages including French (Bergeron et al., 2019), Spanish (Rivas et al., 2021), and Hebrew (Taitelbaum-Swead et al., 2022).

Mandarin Chinese is the most spoken first language in the world, followed by Spanish and English (Figure 1). While both Spanish and English are spoken in geographically dispersed countries, Mandarin Chinese is, for the most part, concentrated in China. It is also a major tonal language (Xu & Zhou, 2011). According to a population-based survey, the standardized prevalence rate of hearing loss was estimated at 15.84% in China (Hu et al., 2016) and China has become one of the most fast-growing markets for CIs. Since the first multichannel CI devices were implanted in

China in 1995, the number of CI recipients has increased exponentially. At present, there are more than 100,000 CI recipients in China—approximately 15% of whom are adults. Given the increasing number of Mandarin-speaking CI patients, there is a need for standardized and validated speech recognition tests in Mandarin Chinese. Responding to the need, researchers and clinicians have developed a variety of materials. Validated speech recognition tests in Mandarin Chinese available for adult listeners vary in speech stimulus type from words (e.g., Han et al., 2009; Ji et al., 2011a, 2011b; Wang et al., 2007; Xi et al., 2010) to sentences (e.g., Fu et al., 2011; Hu et al., 2018; Wong et al., 2007).

Compared to word recognition tests, sentence materials tend to be more “natural” due to their resemblance to daily conversations, involving coarticulation and prosody. Thus, they are more likely to reflect listeners’ speech comprehension in daily life. The first attempt toward standardized sentence test materials for Mandarin-speaking listeners was made by Wong et al. (2007). Following the same rationale as the English HINT (Nilsson et al., 1994), they developed the Mandarin Hearing in Noise Test (MHINT), which consists of 12, 20-sentence lists with low intra-list performance variability and high inter-list reliability. Normative data were available for measuring the reception threshold for sentences in quiet and in noise (Wong et al., 2007). In the meantime, Fu et al. (2011) applied acoustic simulations of CI processing when designing their Mandarin test materials for CI users, namely the Mandarin Speech Perception (MSP) sentence test. They constructed 10 phonetically balanced sentence lists, each containing 10, 7-word sentences. Xi et al. (2012) established a corpus of Mandarin BKB-like sentences with four-talker babble as competing noise, with the homogeneity optimized via psychometric evaluation (HOPE). The sentences can be used in testing both children and adult CI users. More recently, Hu et al. (2018) developed a matrix type of sentence test, the Mandarin Chinese matrix (CMNmatrix) sentence test, for speech recognition measurements in noise with a set of semantically unpredictable and syntactically fixed sentences.

Wang et al. (2015) tested speech recognition in 32 postlingually-deafened CI recipients using the MHINT and Mandarin BKB sentences. The results showed that 4 and 14 out of 32 participants scored 100% correct in quiet with the MHINT and Mandarin BKB sentences, respectively. More than 40% of the participants scored above 85% correct in both tests (Wang et al., 2015). Using the MSP sentences (Fu et al., 2011), Li et al. (2017) also reported that the mean performance for the top one-third of their 35 adult CI users reached >90% correct in quiet. Thus, the ceiling effect in sentence recognition tests is a widespread problem. Such a problem has led to the development of new sentence recognition materials with a higher level of difficulties.



**Figure 1.** The top 10 most-spoken languages (upper bars) and most-spoken first languages in the world (The World Factbook, 2022). The symbols indicate the languages in which AzBio sentence materials have been developed. Hebrew is not shown in the upper bars and French and Hebrew are not shown in the lower bars because they are not in the top 10 categories.

In the present study, we followed the methodology of English AzBio sentence test development and developed a set of Mandarin Chinese sentence lists with a higher level of difficulties. The Mandarin Chinese version of AzBio sentence test we developed here is named CMnBio to follow the similar style of the French (FrBio) and Hebrew (HeBio) versions. Thirty equivalent sentence lists were constructed based on the vocoder-processed sentence recognition tests in the normal-hearing adult listeners. We further validated the use of the CMnBio test in a group of adult CI users. After removing 4 lists that were either too difficult or too easy, we determined that the final version of the CMnBio sentences included 26 highly equivalent lists of 20 sentences.

## Stage I: CMnBio Sentence Construction and List Selection

### Methods

**Sentence construction.** A total of 1,850 Mandarin Chinese sentences were constructed. All sentences were selected from present-day television programs and social media on up-to-date topics, and evaluated to have similar amount of semantic information compared to the original English AzBio sentences by Mandarin-speaking linguists with high proficiency in English. Sentence length was 7 to 15 Chinese characters with 4 to 9 keywords (mean = 6.5, SD = 1). Sentences containing proper nouns and idioms were

excluded. No other restrictions on phonemic composition of the words, vocabulary, or sentence structure were applied to this stage of sentence selection.

**Sentence recordings.** Following the recording procedures of the English AzBio sentence development (Spahr et al., 2012), we recorded all 1,850 sentences for possible inclusion in the CMnBio corpus. Four adult native-Mandarin speakers, two males and two females (aged 20–23 years old), were recruited to read the sentences. One male speaker read 350 sentences and the other three read 500 sentences each. All speakers were undergraduate students professionally trained for public broadcasting in Beijing, China.

The recording took place in a sound-proof booth (noise floor  $\leq 20$  dBA) using a sound level meter (Brüel & Kjær 2250) equipped with a condenser microphone connected to a Creative Audigy™ soundcard. All recordings were made at a sampling frequency of 44.1 kHz with a 16-bit resolution and live-monitored by an examiner using Adobe Audition 3.0. The microphone was positioned approximately 30 cm from the speaker's mouth. All speakers were instructed to read at a normal conversational pace and volume with natural intonation and accentuation, and to avoid excessive enunciation. In the case of mispronunciations, misread words, or any unintended interruption, the speaker was asked to repeat the sentence.

Four audiologists with clinical experience in speech audiometry checked the recorded sentences and excluded

sentences with poor sound quality or low naturalness. Finally, 1,020 out of the 1,850 sentences (i.e., 255 sentences from each speaker) were included in the CMnBio corpus. An additional 80 sentences were included to construct the practice lists. The mean speaking rate across talkers ranged from 4.7 to 5.7 syllables per second, similar to that of the original English AzBio sentences. All 1,020 recorded sentences were normalized using Adobe Audition 3.0 to reach an equal root-mean-square (RMS) level. A calibration signal of the same RMS level was also generated.

**Sentence recognition.** We created 51 sentence lists out of the 1,020 sentences, with each list containing 20 sentences (5 sentences from each talker). In addition, four practice lists of 20 sentences were also constructed. All sentences were processed through a five-channel, noise-excited, vocoder CI simulation in AngelSim™ (Fu, 2010). The overall bandwidth of the noise vocoder was 200 to 7,000 Hz. The spacing of the five bands were determined based on the Greenwood formula (Greenwood, 1990). The lowpass cutoff frequency of the envelope extractor was set at 160 Hz. Our pilot study demonstrated that the 5-channel noise vocoder processing could effectively avoid ceiling and floor effects for the sentence recognition test in adult Mandarin-speaking listeners with normal hearing.

Seventeen native Mandarin-speaking adult listeners (9 males and 8 females; aged  $25.6 \pm 1.4$  years old) were recruited for a listening experiment. Each participant went through an otoscopy exam and puretone audiometry with normal hearing results (i.e., thresholds  $\leq 10$  dB HL between 250 and 4,000 Hz). The vocoder-processed sentences were presented via a laptop connected to a GSI 61 audiometer at 70 dB SPL through the ER-3A insert earphones. Participants were seated in a sound-treated booth and instructed to repeat each sentence they heard. The presentation order of the lists for each subject was designed using Latin squares. The test for each participant was carried out in three sessions on three consecutive days, with 17 lists per session to minimize fatigue effects and practice effects. Each session lasted approximately 45 to 60 min. Prior to each formal testing session, four practice lists were presented to participants to familiarize them with the vocoder-processed signals. The use of human subjects in the present study was reviewed and approved by the Institutional Review Board of the Chinese PLA General Hospital, Beijing, China.

## Results

The mean percent correct score of the 17 normal-hearing participants for all 1,020 sentences was 77.0% correct (SD = 15.1%) with a range from 9.2 to 100% correct (Figure 2). One hundred and fifty sentences, the mean score of which fell within the mean  $\pm 1$  SD of the distribution (i.e., 61.9–92.2% correct), were selected from each talker. The selected

600 sentences had a mean recognition score of 78.0% correct. The mean score for the four talkers (M1, F1, M2, and F2) were 78.9, 78.4, 78.3, and 76.5% correct, respectively. These 600 sentences were used for further validation.

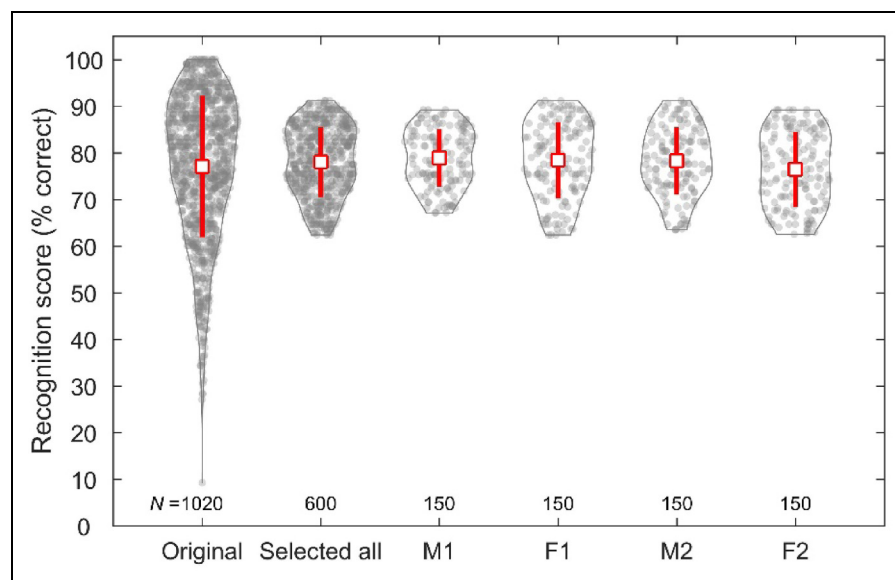
Consistent with the original English version of AzBio sentence test, the list length was set at 20 sentences. Each of the four talkers contributed five sentences to a list. To assign the 150 sentences from each talker to the 30 lists, the 150 sentences were ranked by mean percent correct scores and then sequentially assigned to 30 lists first in descending order (i.e., 30, 29, 28, ..., 1), and then in ascending order (i.e., 1, 2, 3, ..., 30), repeatedly until five sentences of the talker were assigned to each list. This assignment resulted in 30, 20-sentence lists. The mean percent correct scores of the 30 lists ranged from 77.3 to 78.8% correct (Figure 3) with an overall mean of 78.0% correct. A one-way repeated measures analysis of variance (ANOVA) indicated no statistically significant differences in the mean scores among the 30 lists ( $F_{29, 464} = 0.306, p > .05$ ). The final 30 lists had an average of 130.4 keywords (SD = 4, range = 122–143).

## Stage II: CMnBio Sentence List Equivalency Validation with Adult CI Listeners

### Methods

**Participants.** Thirty native Mandarin-speaking adult CI users (10 males and 20 females; aged  $32.1 \pm 12.7$  years old) were recruited for validation of the equivalency and evaluation of the inherent variability in the 30 sentence lists. All participants were diagnosed with bilateral severe to profound sensorineural hearing loss and received cochlear implantation at least 12 months prior to the experiment. Participants were prescreened with a Mandarin monosyllabic word recognition test (Xi et al., 2010) and those who scored  $\geq 40\%$  correct were included in the study. The 30 CI subjects had monosyllabic word scores of 44 to 96% correct (mean = 65.5% correct, SD = 13.9%).

**Procedures.** A pilot study in which recordings were presented in quiet and in +10 dB SNR revealed that few of the CI recipients were able to complete the speech recognition tests in noise. Therefore, the following experiment was conducted only in quiet. Sentences were presented at 65 dB SPL from a loudspeaker located at 0° azimuth on the horizontal plane and 1 meter from the participant in a sound-treated booth. The sentence stimuli were calibrated using a sound level meter (Brüel & Kjær 2250). The participants were instructed to repeat every word of the sentences. Each participant completed three practice lists of 20 sentences that were not included in the 30 test lists prior to formal testing. A break period of 15 min or longer was provided after the completion of every 10 test lists. The test order was counterbalanced across participants in order to eliminate potential learning effects and fatigue effects (see Supplemental Digital



**Figure 2.** Violin plots of the mean sentence recognition scores for 17 normal-hearing listeners. The leftmost violin plot shows the data for the original 1,020 sentences. All sentences were processed through a 5-channel noise vocoder. The next violin plot shows the data for the selected 600 sentences. The next four violin plots are for the four talkers, 150 sentences each. Each violin plot shows the probability density of the data. Each gray dot represents the mean recognition score of one sentence. The squares and the thick vertical lines are the mean  $\pm$  1 SD.

Content 1 for test order for each participant). Each sentence was scored as the number of keywords repeated correctly, and a percent correct score was calculated for each list. The entire test session lasted approximately 3 h with break periods included.

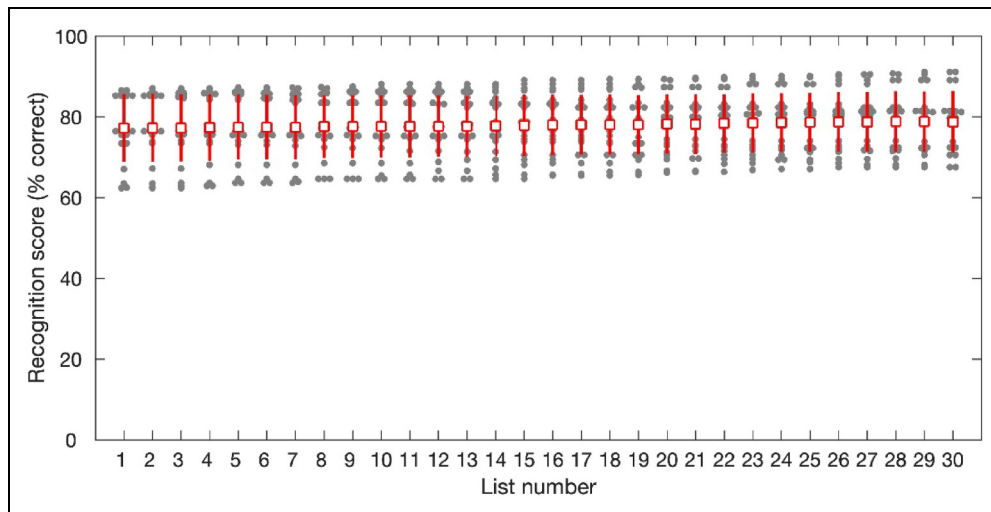
## Results

**Validation.** All 30 CI participants were able to complete the test. The mean scores for the 30 lists achieved by individual CI users ranged from 43.8 to 89.3% correct (mean = 65.0% correct, SD = 13.0%). The distribution of list scores for all 30 CI participants is shown in Figure 4. Only 2 out of the 30 CI participants (6.67%) had a mean score >85% correct. The range of the recognition performance across the 30 lists was between 9.3 and 36.9 percentage points. The standard deviation of the recognition performance across the 30 lists (i.e., the length of the vertical line above or below the mean in Figure 4) was 2.7 to 8.8 percentage points.

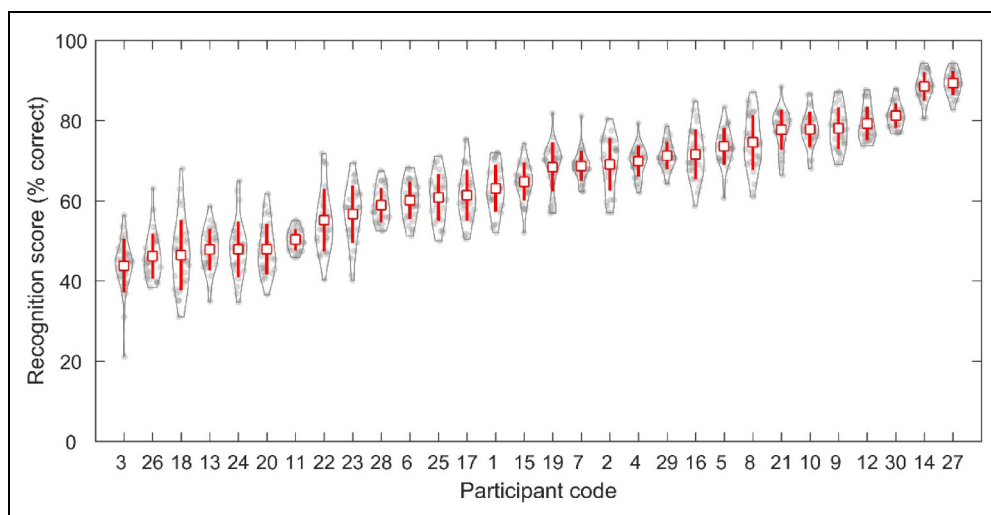
**List variability.** For the 30 lists, the mean recognition performance across all 30 CI participants ranged from 61.7% to 69.0% correct (with an overall mean of 65.0% correct, SD = 1.7%). In order to evaluate variability of recognition scores across the 30 lists without being influenced by the variable performance levels among the CI participants, the list scores of each CI listener were normalized against the mean scores across all 30 lists for that particular CI listener, that is, the normalized score = list score - mean score. This normalization, in essence, removed the effects of various performance levels of the CI participants. Before the

normalization described above was carried out, a rationalized arcsine unit (RAU) transform (Sherbecoe & Studebaker, 2004; Studebaker, 1985) was performed to all percent correct scores. This operation was necessary because our CI participants performed at various levels in sentence recognition and the subsequent statistical analysis would benefit from a more homogenous distribution of variance in the data.

The distribution of the normalized RAU scores for the 30 lists is shown in Figure 5. A one-way repeated measures ANOVA was performed on the normalized RAU scores to examine the equivalency of the 30 lists. A significant main effect of list number ( $F_{29, 841} = 3.08, p < .0001$ ) revealed variability in level of difficulties across the 30 lists. A post hoc Tukey's test showed that 4 of the 30 lists were significantly different from at least one other list. Although the differences were small, Lists 4 and 29 were identified as significantly more difficult and Lists 5 and 10 were identified as easier than the other lists. Excluding those four lists and running the one-way repeated-measures ANOVA for the remaining 26 lists, we found significant differences among them ( $F_{25, 725} = 1.68, p = .02$ ). However, the post hoc Tukey's test found that no lists were significantly different from any of other lists. The mean recognition scores for the 26 lists were from 62.6 to 67.0% correct (SD = 1.2%). Therefore, we determined that the final CMnBio sentence corpus should include these 26 lists (see Supplemental Digital Content 2 for a complete list of all 520 test sentences and 80 practice sentences). The number of Chinese characters in a sentence was 11.9 on average, ranging from 9 to 16.



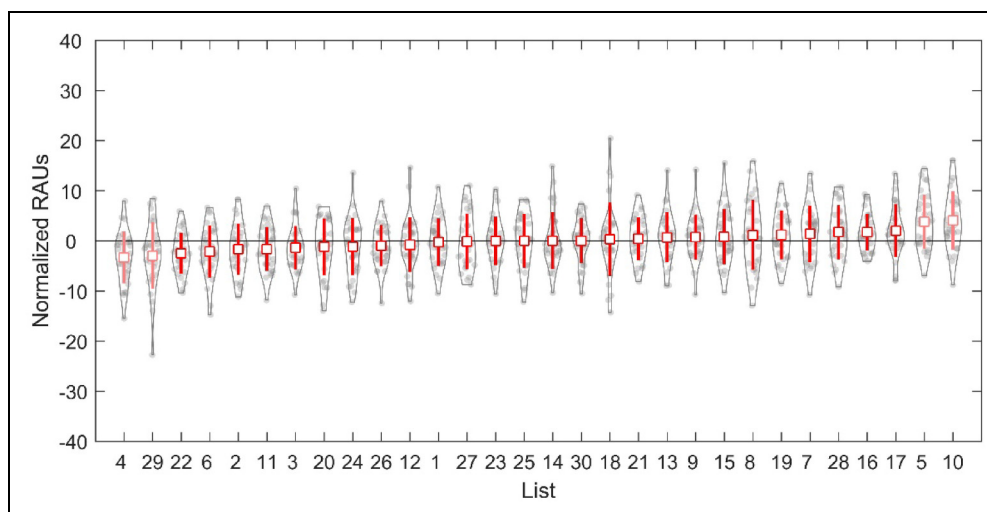
**Figure 3.** Sentence recognition scores for the 30 sentence lists. Each gray dot represents the mean percent correct score of each vocoder-processed sentence averaged across the 17 normal-hearing listeners. The square and the thick vertical lines indicate the mean  $\pm$  1 SD for each list.



**Figure 4.** Sentence recognition scores for 30 cMnBio sentence lists in 30 adult CI participants. The participants were sorted from low to high based on their mean recognition scores. Each violin plot shows the probability density of the data. Each gray dot represents the percent correct score of one sentence list. The square and the thick vertical lines indicate the mean  $\pm$  1 SD for each CI participant.

*Modeling with the binomial distribution.* According to the binomial distribution model of Thornton and Raffin (1978), the variability of speech recognition performance (indicated by SD) is a function of both the starting level of performance and the number of items in the task. The variability is the largest for the mid-range performance and the smallest at the extreme performance. The larger the number of test items in the speech recognition test, the smaller the variability in performance. Spahr et al. (2012) constructed reference ranges (i.e., mean  $\pm$  1.96 SD) that predicted 95% of the distribution of the SD data of the sentence recognition performance. Results based

on visual inspections showed that the reference range constructed with an item number of 40 covered the SD data obtained from the majority of their 15 CI participants. The authors contended that there were 40 independent linguistic units per list for the English AzBio sentences (Spahr et al., 2012). Similar results were reported for the Spanish AzBio sentences (Rivas et al., 2021). We took the same approach as in Spahr et al. (2012) and Rivas et al. (2021) and plotted the SD obtained from the 30 CI participants over the reference range constructed with an item number of 40 based on binomial distribution. The results are shown in Figure 6 (left panel). In the figure,



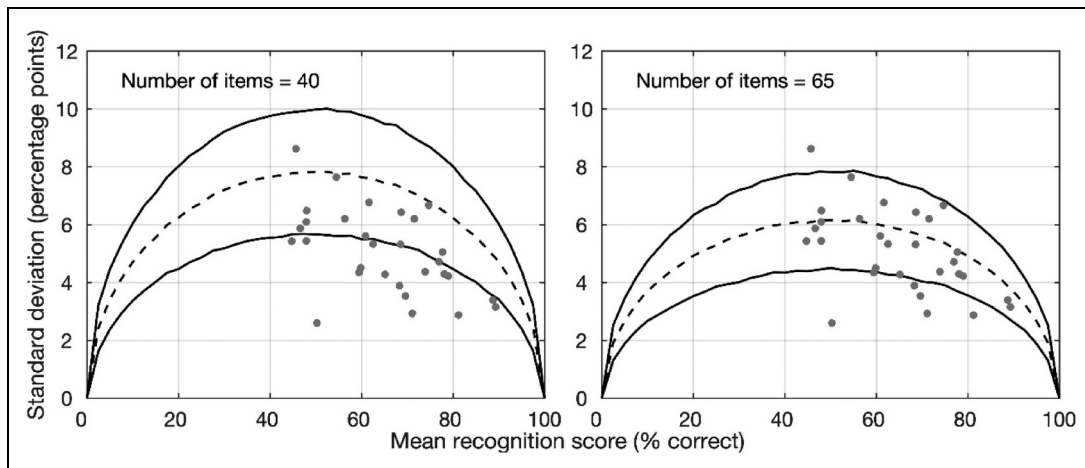
**Figure 5.** Normalized RAU scores for all 30 CMnBio sentence lists. The lists were sorted from low to high based on their mean normalized RAU scores of the 30 CI participants. Each violin plot shows the probability density of the data. Each gray dot represents the percent correct score of one CI participant. The square and the thick vertical lines indicate the mean  $\pm$  1 SD for each sentence list.

the dashed line was the mean and the solid lines were  $\pm 1.96$  SD of the SD data obtained using the bootstrap method. This provided the reference range of 95% of the SD data. Each symbol represents one CI participant's mean and SD of the recognition score across the final 26 lists of the CMnBio sentences. This result indicates that the variability across the CMnBio sentence lists were small. Approximate 2/3 of the data from our 30 CI participants were below the reference range predicted by binomial distribution with an item number of 40. Therefore, the item number of 40 did not fit the CMnBio sentence data. We further increased the number of items in the binomial model. Our visual inspections indicated that when the number of items was between 65 and 80, the binomial prediction fit the data reasonably well with only a few data point outside of the reference range. Figure 6 (right panel) plots the 95% reference range of the binomial model with an item number of 65 along with the recognition data from the 30 CI participants. Next, we chose a different approach to identify the effective number of items in CMnBio sentence lists.

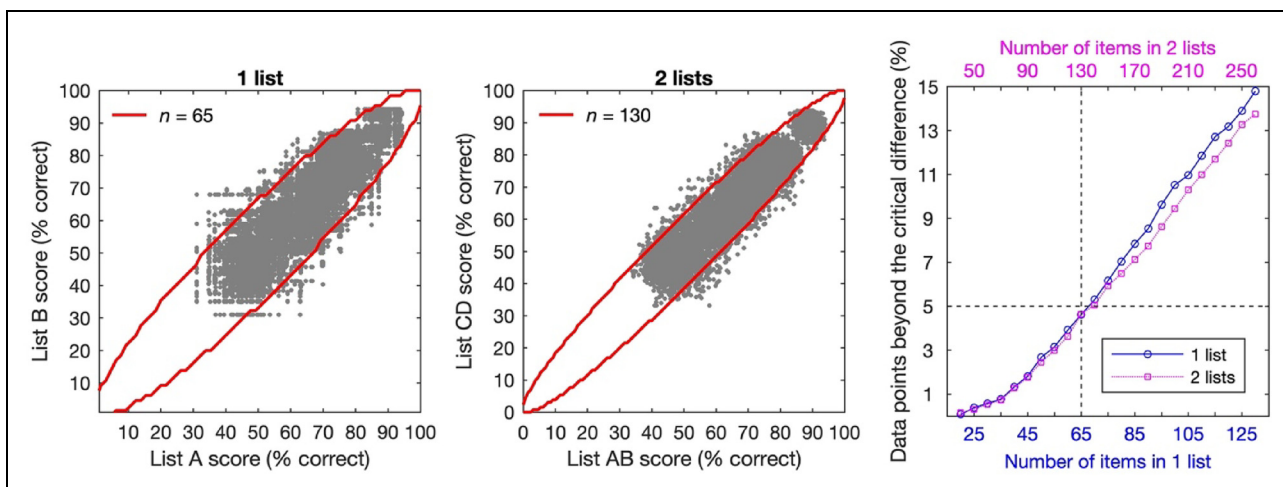
Based on the sentence recognition scores of the final 26 lists by our 30 CI participants, we first created the scatter plot of the scores of any pairs of lists (Figure 7, left panel). For the 26 lists, we have 325 possible pairs. Thus, the scatter plot has 9,750 data points (i.e.,  $325 \times 30$  CI participants). The orders of the lists in the pairs were randomized. If we tested two sentence lists at a time (i.e., one test condition), for each participant with 26 lists, there would be 14,950 possible pairs of 2 lists. To be consistent with the 1-list condition, we randomly chose 325 pairs of scores of 2 lists. Figure 7 (middle panel) shows the scatter plot of the 9,750 data points (i.e.,  $325 \times 30$  CI participants) in the 2-list condition. Next, we calculated the lower and upper limits of the

95% critical difference for percent-correct scores based on Thornton and Raffin (1978) binomial model. The critical differences were dependent on the number of items ( $n$ ) in a test. For the CMnBio sentence corpus, we had 20 sentences per list and an average 130 keywords per lists. Thus, we plotted the lower and upper limits of the critical differences using  $n$  of 20 to 130 in steps of 5 onto the scatter plot of the 1-list condition (Figure 7, left panel,  $n=65$  is shown) and those using  $n$  of 40 to 260 in steps of 10 onto the scatter plot of the 2-list condition (Figure 7, middle panel,  $n=130$  is shown). Finally, we computed the percentage of data points that fell outside the critical differences for each  $n$ . The results are plotted in Figure 7 (right panel). In this panel, the data for the 2-list condition are shifted to align with those for the 1-list condition. We then looked for the number of items that produced approximately 5% of data points beyond the critical difference limits. This analysis indicated that the binomial model (Thornton & Raffin, 1978) with  $n$  of 65 for the 1-list condition and 130 for the 2-list condition best described the variability of the CMnBio sentences.

Table 1 lists the lower and upper limits of the 95% critical differences based on Thornton and Raffin (1978) binomial model. The step size of the scores in this table is 5 percentage points. Supplemental Digital Content 3 provides the critical differences for finer step sizes of the scores. This table is helpful to clinicians who use the CMnBio sentences to test sentence recognition performance in clinical populations. For example, if the test score with one list is 50% correct, then a test score with a different list that falls between 33 and 67% correct (i.e., <17 percentage-point differences) is not statistically significant. However, if the clinician uses two lists for a sentence recognition test and the patient scores 50% correct, a 12 percentage-point difference (i.e.,



**Figure 6.** Variability (SD) as a function of the mean percent correct scores. The predicted SD (dashed line) and 95% reference range (solid lines) were obtained using a bootstrap method based on binomial distribution. The number of items is 40 (left panel) or 65 (right panel). In both panels, each symbol represents one CI participant's mean and SD of the recognition score across the final 26 lists of the CMnBio sentences.



**Figure 7.** List variability and binomial distribution modeling. Left: Scatter plot of recognition scores of any possible pairs of lists from the final 26 CMnBio sentence lists in the 30 CI participants (i.e.,  $325 \times 30$  CI participants). The line represents the lower and upper limits of 95% critical differences based on Thornton and Raffin (1978) binomial model with an item number  $n = 65$ . Middle: Scatter plot of recognition scores of 9,750 randomly chosen pairs of 2 lists from the final 26 CMnBio sentence lists in the 30 CI participants (i.e.,  $325 \times 30$  CI participants). The line represents the lower and upper limits of 95% critical differences based on Thornton and Raffin (1978) binomial model with an item number  $n = 130$ . Right: Percentage of data points that fell outside the critical differences as a function of number of items. The data for the 2-list condition is shifted to the left to align with those of the 1-list condition, using the upper x axis. Note that  $n$  of 65 in the 1-list condition and 130 in the 2-list condition produced approximately 5% of data points that were beyond the critical difference limits (dashed lines).

<38% or >62% correct) will be considered significantly different.

## Discussion

The purpose of this study was to develop and validate a Mandarin Chinese version of the AzBio sentence test. The

new sentence materials should possess a greater level of difficulties to avoid the ceiling effects shown in many high-performing CI recipients when using the traditional sentence materials in quiet. To achieve a relatively high level of difficulties, the speech stimuli were recorded by four different speakers (two males and two females) in a natural conversational manner rather than in an overly enunciated style as



**Table 1.** Lower and Upper Limits of the 95% Critical Differences Predicted by the Binomial Distribution Model.

Score (%)	One list per condition (n = 65)		Two lists per condition (n = 130)	
	Lower	Upper	Lower	Upper
0	0	5	0	2
5	0	15	1	12
10	2	22	4	18
15	6	29	8	24
20	9	35	12	30
25	13	40	16	36
30	16	45	20	42
35	20	52	24	47
40	25	57	28	52
45	30	62	33	57
50	33	67	38	62
55	38	70	43	67
60	43	75	48	72
65	48	80	53	76
70	54	84	58	80
75	59	87	64	84
80	65	91	70	88
85	71	94	76	92
90	78	98	82	96
95	85	99	88	99
100	95	100	98	100

most other materials. In the development of the CMnBio sentences, we strived to achieve a similar level of semantic information to the original English AzBio sentence. Also, the complexity of the sentence structure and demand on vocabulary were kept higher than the Mandarin BKB sentences (Xi et al., 2012). Compared to the MHINT sentences (Wong et al., 2007), we do not have a detailed linguistic analysis. However, there is an apparent difference in the length of the sentences. The MHINT sentences are all 10-character long whereas the lengths of the CMnBio sentences are typically longer than 10 characters. The elevated level of difficulties in the CMnBio sentences was evident in the recognition scores by the normal-hearing adult listeners under the 5-channel noise vocoder conditions. The overall mean recognition score of the 30 lists was 78.0% correct (Figure 3). Our previous research has shown that under similar noise vocoder conditions, normal-hearing adult listeners could easily achieve >90% correct with the Mandarin BKB sentence materials (Xi et al., 2019). In a recent study, we compared sentence recognition of the MHINT and the CMnBio sentences under various channel conditions of vocoder processing (Xu et al., 2021). The recognition scores were 93.4% and 73.9% correct, respectively, for the two types of sentence materials under 5-channel noise vocoder conditions. In the present study, only 2 out of the 30 CI participants scored between 85 to 90% correct (Figure 4). Note that all the CI participants had at least moderately good speech recognition

(i.e., >40% correct word recognition) when enrolled in the study. Clearly, a higher level of difficulties in the CMnBio sentences has been achieved and the ceiling effects in speech recognition tests can be minimized with the new CMnBio sentence test.

The level of difficulties of the CMnBio sentences developed in this study appeared to be slightly higher than that of other languages developed earlier. Our 17 normal-hearing listeners scored 78.0% correct using the 5-channel noise vocoder CI simulation. Using a similar 5-channel vocoder CI simulations, the normal-hearing listeners scored 85.2% correct with the English AzBio sentences (Spahr et al., 2012), 85.0% correct with the Spanish AzBio sentences (Rivas et al., 2021), and 82.0% correct with the Hebrew AzBio sentences (Taitelbaum-Swead et al., 2022), respectively. In a previous study, we compared recognition of the English AzBio and CMnBio sentences under various vocoder conditions. We found that the CMnBio sentences produced recognition scores that were approximately 9 to 12 percentage points lower than the English AzBio sentences (Xu et al., 2021). The French version of AzBio sentences seemed to have the lowest level of difficulties among all languages in which AzBio sentences have been developed. Bergeron et al. (2019) stated that all their French-speaking normal-hearing participants performed at 100% correct with the 5-channel CI simulations. They chose to use 4-channel CI simulations instead and found that the mean recognition score was approximately 85% correct with the 4-channel CI simulations. Therefore, while the English, Spanish, and Hebrew versions of AzBio sentences showed a very similar level of difficulties, the Mandarin Chinese and French versions appeared to diverge from them. We should keep in mind such a divergence in the level of difficulties when carrying out cross-language comparison of CI performance. It is worth noting that it might be unrealistic to create sentence materials with identical level of difficulties among different languages for a variety of linguistic reasons. Various versions of AzBio sentences created so far have achieved the desired goals with elevated level of difficulties compared to other sentence materials.

Another feature of the CMnBio sentence corpus is that it contains a larger number of lists. Our final corpus has 26 test lists and 4 practice lists (see Supplemental Digital Content 2). This number is slightly smaller than those in English (33 lists, Spahr et al., 2012), French (30 lists, Bergeron et al., 2019), Spanish (42 lists, Rivas et al., 2021), and Hebrew (33 lists, Taitelbaum-Swead et al., 2022) but is much greater than that of the Mandarin HINT sentences (12 lists) (Wong et al., 2007). This feature makes it possible to conduct large-scale studies involving multiple experimental conditions without repeated use of materials. It also provides clinicians with an opportunity to track changes in patients' performance over time or across conditions. For these reasons, we anticipate that the CMnBio sentences would be

used to assess speech perception of adult CI users in the clinics as well as in the research laboratories.

The validation process showed that the CMnBio sentences possessed satisfactory inter-list reliability. For the 30 original lists, the mean recognition performance across all 30 CI participants ranged from 62 to 69% correct (range = 7 percentage points,  $SD = 1.7$ ). In comparison, Spahr et al. (2012) showed that the average scores for the 33 English AzBio sentence lists in their 15 CI listeners ranged from 62 to 79% correct (range = 17 percentage points,  $SD = 3.8$ ). The small inter-list variability of the CMnBio sentences is desirable because in clinical practice, it is often necessary to compare speech recognition performance among different conditions or at different time points using different sentence lists.

The small inter-list variability was also reflected in the relatively small standard deviation of sentence recognition scores across the final 26 lists in the 30 CI participants (Figure 6). Using the binomial distribution modeling (Thornton & Raffin, 1978), a number of items of 40 that fitted the English and Spanish AzBio sentence materials (Rivas et al., 2021; Spahr et al., 2012) or 50 that fitted the FrBio sentences (Bergeron et al., 2019) did not fit our CMnBio sentence materials. As we increased the number of items to a range of 65 to 80, the binomial distribution model fitted our data reasonably well by visual inspection. We then used a more quantitative approach to assess the number of items in a sentence list and found an item number of 65 in a sentence list best fitted our sentence materials (Figure 7). With an approximately 130 keywords in a sentence list in the CMnBio sentence materials, it appears that the number of independent items was one half of the number of keywords. Keidser et al. (2002) also suggested that for a BKB English sentence list containing 50 keywords, an item number of 25 would be an appropriate estimate. Our estimated number of items (65) is greater than those in either English or Spanish AzBio sentence materials (40). Note that the average number of words in English or Spanish is 142 and the scoring is based on words not keywords. Another important difference was the methods used to derive the lower and upper limits between the present study and Spahr et al. (2012) or Rivas et al. (2021). In the present study, we derived the lower and upper limits of the 95% critical differences using exactly the methods and terminology of Thornton and Raffin (1978). However, Spahr et al. (2012) and Rivas et al. (2021) apparently used the normal approximation method of the binomial confidence interval to derive the lower and upper limits of the 95% confidence interval. In essence, the lower and upper limits that we derived with numbers of items of 65 and 130 (Table 1) were quite similar to their derived limits with numbers of items of 40 and 80, respectively [Table 1 of Spahr et al. (2012) and Rivas et al. (2021)]. Thus, the number of items might not be different among the studies after all and might not imply that the contexture information among words in the AzBio

sentences are different among languages. Clinically, the exact number of items may not be important as long as the derived lower and upper limits are accurate. For the present study, such limits were calculated based on the data from 30 adult CI participants with reasonably good speech recognition performance. A large scale of empirical data with a diverse CI population would be required to establish the most reliable limits of the critical differences for future clinical uses.

## Conclusions

The CMnBio sentence test has been developed for speakers of Mandarin Chinese. The test was created using a methodology similar to that used in the development of the English AzBio sentence test and the reliability and validity were confirmed in adult CI recipients. The new test has 26 lists of 20 sentences each. The lists have equal intelligibility. Compared to similar sentence tests in English, Spanish, French, and Hebrew, the Mandarin Chinese material appears to be slightly more difficult. This elevated level of difficulties should serve the purpose of avoiding ceiling effects when testing sentence recognition in Mandarin-speaking CI users. Finally, the lower and upper limits of the 95% critical differences were derived from the modeling with binomial distribution and they should provide guidance for clinical detection of a significant difference in recognition scores.

## Acknowledgments

The authors express gratitude to Drs. Rene Gifford, Tony Spahr, and Kevin Yuen for their constructive suggestions and comments on an earlier version of this manuscript. The authors also thank Feng Lu, Ying Fu, Wei Wang, and Meng-Di Hong for their assistance in participant recruitment, and Victoria Costa for her editorial assistance in the preparation of the article.

## Author Note

Portions of this article were previously presented during the 10th Asia-Pacific Symposium on Cochlear Implants and Related Sciences (APSCI 2015), April 30–May 3, 2015.

## Author Contribution

X. X. and L. X. designed the experiments. X. X., Y. W., Y. S., R. G., X. Q., and Q. W. performed the experiments. L. X., X. X., Y. W., and S. L. analyzed the data and wrote the main article. All authors discussed the results and implications and commented on the article at all stages.

## Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.


## Funding


The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work

was supported by the National Natural Science Foundation of China, National Key R&D Program of China (grant number 61370023, 81460099, 2020YFC2004005).

### ORCID iDs

Xin Xi  <https://orcid.org/0000-0001-7806-8396>

Siqi Li  <https://orcid.org/0000-0001-8660-6414>

Li Xu  <https://orcid.org/0000-0002-0988-7934>

### Supplemental Material

Supplementary material for this article is available online

### References

- Bassim, M. K., Buss, E., Clark, M. S., Kolln, K. A., Pillsbury, C. H., Pillsbury, H. C., III, & Buchman, C. A. (2005). MED-EL Combi40+ cochlear implantation in adults. *The Laryngoscope*, *115*, 1568–1573. <https://doi.org/10.1097/01.mlg.0000171023.72680.95>
- Bergeron, F., Berland, A., Fitzpatrick, E. M., Vincent, C., Giasson, A., Leung Kam, K., Chafiq, W., Fanouillère, T., & Demers, D. (2019). Development and validation of the FrBio, an international French adaptation of the AzBio sentence lists. *International Journal of Audiology*, *58*, 510–515. <https://doi.org/10.1080/14992027.2019.1581950>
- Boothroyd, A., Hanin, L., & Hnath, T. (1985). *A sentence test of speech perception: Reliability, set equivalence, and short term learning*. Internal Report RCI 10. Speech and Hearing Sciences Research Center, City University of New York.
- Fu, Q. J. (2010). AngelSim™ (Version 1.05.03) [Software]. [http://angelsim.emilyfufoundation.org/angelsim\\_about.html](http://angelsim.emilyfufoundation.org/angelsim_about.html)
- Fu, Q. J., Zhu, M., & Wang, X. (2011). Development and validation of the Mandarin speech perception test. *The Journal of the Acoustical Society of America*, *129*, 267–273. <https://doi.org/10.1121/1.3590739>
- Gifford, R. H., Shallop, J. K., & Peterson, A. (2008). Speech recognition materials and ceiling effects: Considerations for cochlear implant programs. *Audiology and Neuro-Otology*, *13*, 193–205. <https://doi.org/10.1159/000113510>
- Greenwood, D. D. (1990). A cochlear frequency-position function for several species—29 years later. *The Journal of the Acoustical Society of America*, *87*, 2592–2605. <https://doi.org/10.1121/1.399052>
- Han, D., Wang, S., Zhang, H., Chen, J., Jiang, W., Mannell, R., Newall, P., & Zhang, L. (2009). Development of Mandarin monosyllabic speech test materials in China. *International Journal of Audiology*, *48*, 300–311. <https://doi.org/10.1080/14992020802607456>
- Hirsh, I. J., Davis, H., Silverman, S. R., Reynolds, E., Eldert, E., & Benson, R. W. (1952). Development of materials for speech audiometry. *Journal of Speech and Hearing Disorders*, *17*, 321–337. <https://doi.org/10.1044/jshd.1703.321>
- Hu, H., Xi, X., Wong, L., Hochmuth, S., Warzybok, A., & Kollmeier, B. (2018). Construction and evaluation of the Mandarin Chinese matrix (CMNmatrix) sentence test for the assessment of speech recognition in noise. *International Journal of Audiology*, *57*, 838–850. <https://doi.org/10.1080/14992027.2018.1483083>
- Hu, X. Y., Zheng, X. Y., Ma, F. R., Long, M., Han, R., Zhou, L. J., Wang, F., Gong, R., Pan, T., Zhang, S. X., Du, B., Jin, P., Guo, C. Y., Zheng, Y. Q., Liu, M., He, L. H., Qiu, J. H., Xu, M., Song, L., ..., Wang, S. P. (2016). Prevalence of hearing disorders in China: A population-based survey in four provinces of China (in Chinese). *Chinese Journal of Otorhinolaryngology—Head & Neck Surgery*, *51*, 819–825. <http://doi.org/10.3760/cma.j.issn.1673-0860.2016.11.004>
- Ji, F., Xi, X., Chen, A. T., Ying, J., Wang, Q. J., & Yang, S. M. (2011a). Development of a Mandarin monosyllable test material with homogenous items (I): Homogeneity selection. *Acta Otolaryngologica*, *131*, 962–969. <https://doi.org/10.3109/00016489.2011.574646>
- Ji, F., Xi, X., Chen, A. T., Zhao, W. L., Zhang, X., Ni, Y. F., Yang, S. M., & Wang, Q. (2011b). Development of a mandarin monosyllable test material with homogenous items (II): Lists equivalence evaluation. *Acta Otolaryngologica*, *131*, 1051–1060. <https://doi.org/10.3109/00016489.2011.583267>
- Keidser, G., Ching, T., Dillon, H., Agung, K., Brew, C., Brewer, S., Fisher, M., Foster, L., Grant, F., & Storey, L. (2002). The National Acoustic Laboratories' (NAL) CDs of speech and noise for hearing aid evaluation: Normative data and potential applications. *Australian and New Zealand Journal of Audiology*, *24*, 16–35. <https://doi.org/10.1375/audi.24.1.16.31112>
- Li, Y., Wang, S., Su, Q., Galvin, J. J., & Fu, Q.-J. (2017). Validation of list equivalency for Mandarin speech materials to use with cochlear implant listeners. *International Journal of Audiology*, *56*, S31–S40. <https://doi.org/10.1080/14992027.2016.1204564>
- Messersmith, J. J., Entwisle, L., Warren, S., & Scott, M. (2019). Clinical practice guidelines: Cochlear implants. *Journal of American Academy of Audiology*, *30*, 827–844. <https://doi.org/10.3766/jaaa.19088>
- MSTB (2011). The new Minimum Speech Test Battery <http://www.auditorypotential.com/MSTBfiles/MSTBManual2011-06-20%20.pdf>
- Nilsson, M. J., Soli, S., & Sullivan, J. (1994). Development of the Hearing in Noise Test for the measurement of speech reception thresholds in quiet and in noise. *The Journal of the Acoustical Society of America*, *95*, 1085–1099. <https://doi.org/10.1121/1.408469>
- Peterson, G. E., & Lehiste, I. (1962). Revised CNC lists for auditory tests. *Journal of Speech and Hearing Disorders*, *27*, 62–70. <https://doi.org/10.1044/jshd.2701.62>
- Rivas, A., Perkins, E., Rivas, A., Rincon, L. A., Litvak, L., Spahr, T., Dorman, M., Kessler, D., & Gifford, R. (2021). Development and validation of the Spanish AzBio sentence corpus. *Otology & Neurotology*, *42*, 154–158. <https://doi.org/10.1097/MAO.0000000000002970>
- Sherbecoe, R. L., & Studebaker, G. A. (2004). Supplementary formulas and tables for calculating and interconverting speech recognition scores in transformed arcsine units. *International Journal of Audiology*, *43*, 442–448. <https://doi.org/10.1080/14992020400050056>
- Spahr, A. J., Dorman, M. F., Litvak, L. M., Van Wie, S., Gifford, R. H., Loizou, P. C., Loiselle, L. M., Oakes, T., & Cook, S. (2012). Development and validation of the AzBio sentence lists. *Ear and Hearing*, *33*, 112–117. <https://doi.org/10.1097/AUD.0b013e31822c2549>
- Studebaker, G. A. (1985). A “rationalized” arcsine transform. *Journal of Speech and Hearing Research*, *28*, 455–462. <https://doi.org/10.1044/jshr.2803.455>

- Taitelbaum-Swead, R., Dahan, T., Katzenel, U., Dorman, M. F., Litvak, L. M., & Fostic, L. (2022). Azbio sentence test in Hebrew (HeBio): Development, preliminary validation, and the effect of noise. *Cochlear Implant International*, 23, 270–279. <https://doi.org/10.1080/14670100.2022.2083285>
- The World Factbook (2022). People and Society—Languages. <https://www.cia.gov/the-world-factbook/countries/world/#people-and-society>
- Thornton, A. R., & Raffin, M. J. (1978). Speech-discrimination scores modeled as a binomial variable. *Journal of Speech and Hearing Research*, 21, 507–518. <https://doi.org/10.1044/jshr.2103.507>
- Tillman, T. W., & Carhart, R. (1966) An expanded test for speech discrimination utilizing CNC monosyllabic words. Northwestern University Auditory Test No. 6. SAM-TR-66-55. Tech Rep SAM-TR. 1966 Jun:1-12. <https://doi.org/10.21236/ad0639638>
- Wang, S., Mannell, R., Newall, P., Zhang, H., & Han, D. (2007). Development and evaluation of Mandarin disyllabic materials for speech audiometry in China. *International Journal of Audiology*, 46, 719–731. <https://doi.org/10.1080/14992020701558511>
- Wang, Y., Shi, Y., Fu, Y., Wang, Q., Fu, Y., & Xi, X. (2015). Study of ceiling effect of commonly used Chinese recognition materials in post-lingual deafened patients with cochlear implant. *Journal of Clinical Otorhinolaryngology—Head & Neck Surgery*, 29, 298–303. <http://doi.org/10.13201/j.issn.1001-1781.2015.04.003>
- Wong, L. L., Soli, S. D., Liu, S., Han, N., & Huang, M. W. (2007). Development of the Mandarin Hearing in Noise Test (MHINT). *Ear and Hearing*, 28(2 Suppl), 70S–74S. <https://doi.org/10.1097/AUD.0b013e31803154d0>
- Xi, X., Ching, T. Y., Ji, F., Zhao, Y., Li, J. N., Seymour, J., Hong, M. D., Chen, A. T., & Dillon, H. (2012). Development of a corpus of Mandarin sentences in babble with homogeneity optimized via psychometric evaluation. *International Journal of Audiology*, 51, 399–404. <https://doi.org/10.3109/14992027.2011.642011>
- Xi, X., Ji, F., Chen, A. T., Zhao, W. L., Zhao, Y., Xu, J., Qiu, C. Y., Li, J. H., & Han, D. Y. (2010). Development and evaluation of standardized Mandarin monosyllabic audiometric materials. *Chinese Journal of Otorhinolaryngology—Head & Neck Surgery*, 45, 7–13. <http://doi.org/10.3760/cma.j.issn.1673-0860.2010.01.003>
- Xi, X., Patton, A., Gao, R., Qi, B., Xu, X., Xu, S., Fang, Y., & Xu, L. (2019). A study of level of difficulties of commonly-used sentence recognition materials in Mandarin Chinese and American English using vocoder processing. In Presented at China National Annual Otology Conference, Kunming, Yunan, China.
- Xu, L., Xi, X., Patton, A., Wang, X., Qi, B., & Johnson, L. (2021). A cross-language comparison of sentence recognition using American English and Mandarin Chinese HINT and AzBio sentences. *Ear and Hearing*, 42, 405–413. <https://doi.org/10.1097/AUD.0000000000000938>
- Xu, L., & Zhou, N. (2011). Tonal languages and cochlear implants. In F.-G. Zeng, A. N. Popper, & R. R. Fay (Eds.), *Auditory prostheses: New horizons* (pp. 341–364). Springer Science+Business Media, LLC.