List equivalency and critical differences of a Mandarin BKB sentence in babble noise test for adults and pre-school children with normal hearing

Xin Xi[a,b]*, Jia-Nan Li[a,b]*, Kevin C. P. Yuen[c], Ai-Ting Chen[a,b], Si-Qi Li[d], Meng-Di Hong[a,b], Qian Wang[a,b], Fei Ji[a,b], Harvey Dillon[e,f] and Teresa Y. C. Ching[g, h, i].

*These authors have equal contribution to this paper and the test development.

[a] *Department of Otolaryngology - Head & Neck Surgery, the Sixth Medical Center of PLA General Hospital, Beijing, China;*

[b] *National Clinical Research Center for Otolaryngologic Diseases, Beijing, China*;

[c] *Department of Special Education and Counselling, The Education University of Hong Kong, HKSAR, China;*

[d] *School of Communication Science, Beijing Language and Culture University, Beijing, China;*

[e] *Department of Linguistics, Macquarie University, Sydney, Australia;*

[f] *Manchester Centre for Audiology and Deafness, School of Health Sciences, University of Manchester, United Kingdom;*

[g] *Macquarie School of Education, Macquarie University, Sydney, Australia;*

[h] *NextSense Institute, Sydney, Australia;*

[i] *School of Health and Rehabilitation Sciences, University of Queensland, Brisbane, Australia.*

Correspondence:

Teresa Y. C. Ching, Macquarie School of Education, Macquarie University, Sydney 2109, Australia; teresaychingmail.com

Harvey Dillon, Department of Linguistics, Macquarie University, Sydney 2109, Australia; harvey.dillon@mq.edu.au

Abstract

Purpose: To determine the speech recognition equivalence of Mandarin BKB sentence lists with adults and children with normal hearing.

Method: A total of 32 lists each of 9 sentences were compiled from a corpus of BKB-like sentences with paired babble in Mandarin. Inter-list equivalence, critical differences, and sensitivity of performance to signal-to-noise ratio (SNR) were examined. Experiment 1 included 64 native Mandarin-speaking adults with normal hearing. Experiment 2 included 54 native Mandarin-speaking children with normal hearing aged 4–6 years old.

Results: Among the 32 sentence lists, 28 lists were confirmed to be equivalent in adults, with a mean SNR required for 50% correct (SNR_{50}) of -5.9 ± 0.1 dB, a mean slope of 22.3 ± 1.5%/dB, and the grand 95% critical difference (CD) was subsequently calculated as 27.2% for score. From the 28 equivalent lists, 27 lists were selected and observed to be equivalent in children, with a mean SNR_{50} threshold of -2.0 ± 0.2 dB, a mean slope of 15.8 ± 1.1%/dB, and a grand 95% CD of 24.6% for score.

Conclusion: The Mandarin BKB sentences in babble noise test offers an opportunity for clinicians and researchers to assess speech understanding in adults and pre-school children in an efficient manner. For comparisons of performance in different test conditions, 28 equivalent lists are available for adults, and 27 equivalent lists for pre-school children. The 95% critical difference values can be used for total percentage correct or SNR for 50% performance. Future work will examine the clinical utility for school-aged children and children who are deaf and hard of hearing.

Keywords: speech perception; Mandarin sentence test; babble noise

Introduction

Speech audiometry is a fundamental tool in the assessment of hearing disorders and the evaluation of rehabilitative outcomes. It can not only aid in determining the degree and type of hearing loss in conjunction with pure-tone audiometry, but also help make clinical decisions about appropriate gain and maximum output of hearing aids and other auditory prosthetics for hearing-impaired patients, as well as help assess how well they hear in noise.

The primary concern of most hearing-impaired people is the difficulties in understanding speech in noisy environments (Moore, 2003), which cannot be accurately predicted from hearing thresholds (Killion et al., 2004). Therefore, speech-in-noise tests have been recommended as routine practice for clinical evaluation of patients using hearing aids (HAs) and cochlear implants (CIs) (Wilson, 2004). The target speech can range in complexity from nonsense syllables, digits, to monosyllabic words, multisyllabic words, and sentences. Compared with word recognition tests, sentence materials contain redundant information and dynamic amplitude variation that are typical of connected speech. Given its better resemblance to speech situations that users of HAs or CIs would encounter in realistic listening environments, sentence tests may be a more valid measurement of real-world speech intelligibility.

Since HAs and CIs have become increasingly common for the intervention of hearing loss in China, there has been an urgent need for standardized speech recognition test materials. As the official language of China, Mandarin Chinese is spoken by over 80% of its population (China State Language Commission, 2021). The popularity of Mandarin enables the measurement of speech perception among people from different dialect regions. Currently, there are a limited number of sentence recognition tests available in Mandarin Chinese. The first standardized sentence test is the Mandarin Hearing in Noise Test (MHINT) developed by Wong and her colleagues (2007)

following the same rationale as the English HINT (Nilsson, Soli & Sullivan, 1994). The MHINT comprises 24 lists of 10 sentences each, with normative data available for measuring the reception threshold for sentences in quiet as well as in noise. Fu, Zhu and Wang (2011) focused on the differences in tone perception between listeners with normal hearing and CI users and developed a set of phonetically balanced sentence lists named the Mandarin speech perception (MSP) test. Hu et al. (2018) developed a sentence test using semantically unpredictable sentences with fixed grammatical structure to assess speech perception in noise, namely the Mandarin Chinese matrix (CMN matrix) sentence test.

When it comes to pediatric sentence tests, a closed-set sentence identification test (Mandarin Paediatric Speech Intelligibility test, MPSI; Zheng et al., 2009) has been developed to evaluate children's speech recognition first in quiet and then in the presence of a competing sentence at signal-to-noise ratios (SNRs) of +10, +5, 0, –5 and –10 dB. Since there are children scoring at the ceiling of the test, Chen and Wong (2020) created a sentence test with a higher level of difficulty. They selected sentences suitable for evaluating children aged 6 years and above (primary school age in Mainland of China) from the MHINT, and developed an open-set pediatric sentence test – the Mandarin version of the Hearing in Noise Test for Children (MHINT-C). The MHINT-C comprises 12 lists each of 10 sentences, with each sentence consisting of 10 characters. The test was verified for use with Mandarin-speaking children aged 6–18 years with age-specific correction factors established. The ability to understand speech in the presence of competing sounds, particularly speech, is especially important for preschool aged children, who spend much of their waking hours in acoustic environments where noise (speech or otherwise) levels can be high while they learn and acquire language (Newman et al., 2015; Leibold & Buss, 2019). Further work is needed

105 to examine how pre-school children perform in environments such as classrooms, and to
106 devise appropriate tools for evaluating young children's performance using hearing aids
107 and cochlear implants (Litovsky et al., 2006). Given the trend of an increasing number
108 of young candidates for cochlear implantation (Li et al., 2017), it is necessary to devise
109 open-set sentence tests for the evaluation of speech perception in noise abilities in
110 children younger than 6 years of age (Ching et al, 2018) to be used in addition to
111 closed-set sentence tests.

112 In English-speaking countries, the Bamford-Kowal-Bench (BKB) sentences
113 (Bench & Bamford, 1979) are one of the most commonly used speech materials. BKB
114 sentences are short, highly redundant and rich in semantic and syntactic context. So,
115 they are appropriate for assessing the speech recognition of children as well as adults.
116 The BKB sentence test is presented in an open-set format and is commonly used for
117 candidacy selection for CIs, according to FDA-approved criteria. Several speech-in-
118 noise test materials have also been developed based on BKB sentences, such as the
119 Hearing in Noise Test (HINT; Nilsson, Soli & Sullivan, 1994) and the BKB-Sentences
120 in Noise test (BKB-SIN; Etymotic Research, 2005). The BKB-SIN test has been
121 included in the latest version of the Minimum Speech Test Battery for adult CI users
122 (MSTB, 2011).

123 Given the effectiveness and practicality of BKB sentences, they can be a proper
124 reference for standardized Mandarin sentence tests. In line with the principles of the
125 BKB sentence test, Xi and his colleagues (2012) constructed a corpus of Mandarin
126 BKB-like sentences in babble noise with *H*omogeneity *O*ptimized via *P*sychometric
127 *E*valuation, the HOPE corpus. In the study, 647 6- to 8-character sentences within the
128 syntactic and semantic abilities of children were selected from a speech corpus of pre-
129 school children (Liu et al. 2008). A total of 309 sentences were retained based on the

speech-in-noise recognition performance of 96 Mandarin-speaking adults. Following the International Collegium of Rehabilitative Audiology (ICRA) guidelines for the development of multilingual speech tests (Akeroyd et al., 2015), the SNR of each keyword was adjusted by the difference between that keyword's speech recognition threshold (SNR_{50}) and the average SNR_{50} across all 309 sentences, to equalize the difficulty of test items. The homogeneity of adjusted keywords was verified in another group of 64 Mandarin-speaking subjects. Although the design of the test material is appealing for hearing evaluation in young children and research, the psychometric data available on the sentences, to date, were based on adult performance. As the sentences were not compiled into lists, information on list equivalency and critical difference were not available.

The purpose of the current study was to compile a set of Mandarin sentence in babble noise test lists from the homogeneous sentence-babble pairs in the HOPE corpus and determine the list equivalency and critical differences for adults and pre-school children with normal hearing.

Experiment 1: List construction and equivalency in adults with normal hearing

Materials and methods

Materials

A total of 32 sentence lists (Lists A to Z and Lists 1 to 6), each comprising nine scoring sentences and one preliminary example sentence with paired frozen babble noise, were constructed from the HOPE corpus. The target sentences were read by a 20-year-old male native Mandarin speaker, while the babble noise was created from

153 connected discourse read by two males and two females aged from 28 to 38 (Details of
154 the generation of target speech and babble noise can be seen in Xi et al., 2012).
155 The detailed procedures of list compilation were described as follows. First, all
156 309 sentences in the corpus were categorized in terms of sentence type (statements,
157 interrogatives, exclamatory, and imperatives) and ranked according to syntactic
158 difficulty corresponding to the developmental stages of children. According to Zhu and
159 Miao's (1990) theory of child language development, syntactic difficulty of sentences
160 can be categorized into four types – simple sentence without any modifiers, complex
161 predicate sentence, complex subject/object sentence and compound sentence. Syntactic
162 development progressed from acquisition of simple sentences without any modifiers to
163 compound sentences. Next, 288 sentences were selected as scoring sentences to form 32
164 lists based on three rules: (1) sentences belonging to different types and syntactic
165 difficulty should be balanced across lists as much as possible; (2) sentences about the
166 same topic should be distributed as balance as possible across lists; for example,
167 sentences talking about having food would not appear too many times in a single list;
168 (3) the number of keywords for scoring in each list was 50. The remaining sentences
169 from the corpus were then used as the preliminary example sentence in each list, and
170 also in four practice lists.
171 The keyword-based scoring rule and selection of keywords remained the same as
172 the study on corpus construction (Xi et al., 2012) since item homogeneity was verified
173 in terms of keywords. All keywords for scoring were selected from content words and
174 function words that received primary stress in each sentence. Each keyword contained
175 one to three characters. The unit of measurement was defined as word instead of
176 character because single characters in Chinese can be combined to form a word
177 representing a semantic unit. With additional semantic cues, listeners tend to recognize

178 the word as a whole rather than separated single characters. The number of keywords
179 for scoring in each sentence ranged from three to seven, with a total of 50 keywords in
180 each list.

181 All 32 test lists and the four practice lists were burned onto a compact disc (CD).
182 Each sentence, in one channel, was preceded by a 1.6 kHz beep 500 ms before the start
183 of the sentence and was time-locked to a paired segment of babble noise in another
184 channel. The babble noise commenced 2 seconds prior to each sentence and ceased 500
185 ms after the sentence. The inter-stimuli interval was four seconds. The CD also
186 contained a 60-second calibration noise with its spectrum matched with International
187 Long-Term Average Speech Spectrum (ILTASS; Byrne et al, 1994), which
188 corresponded with the stimulus sentences and babble segments that had also been
189 filtered to match the ILTASS (Xi et al., 2012). The calibration noise's level was set at
190 the same RMS (Root of Mean Square) of equivalent continuous level as the sentences
191 and babble segments.

192 *Participants*

193 A total of 64 native Mandarin-speaking adults (42 females and 22 males) with no
194 known history of hearing or speech disorders were recruited to participate. All
195 participants were born and grew up in Beijing or nearby provinces so that the effects of
196 dialect were minimized. They were aged 18–30 years (mean = 23.8 years, SD = 2.8
197 years), with air-conduction pure-tone thresholds ≤15 dB HL at audiometric frequencies
198 from 250 to 8000 Hz as well as normal type A tympanograms. The recruitment of
199 human subjects in the present study was reviewed and approved by the Institutional
200 Review Board of the Chinese PLA General Hospital, Beijing, China.

Experimental design and procedure

The experiment was conducted in a sound-proof booth in the Chinese PLA General Hospital, Beijing. The sentences were presented via a Marantz 5400 CD player to channel 1 of a GSI 16 audiometer at a fixed level of 60 dB SPL. The babble noise was presented to Channel 2 of the audiometer, with its intensity manually adjusted by the experimenter. As the SNR_{50} for adult listeners with normal hearing was found to be approximately -6 dB SNR (SD = 1.1 dB) in the HOPE corpus study (Xi et al., 2012), sentences were tested at -9, -7, -5 and -3 dB SNRs in this experiment. The sentences (Channel 1) and babble noise (Channel 2) were mixed by the audiometer and connected to a GSI loudspeaker 1 m in front of the participants. The sound field was calibrated according to the Sound Field Measurement Tutorial (American Speech-Language-Hearing Association, 1991). A Brüel & Kjaer 2250 sound level meter was used to measure the loudspeaker's output of the calibration noise with its microphone located at the reference point corresponding to the position of the participant's head. Level adjustments for sentences and noise were based on the measured level of calibration signal.

The participants were instructed to repeat as many words in the sentences as possible after each sentence was presented. Performance was scored by the experimenter during the test in terms of number of keywords correctly repeated. Four practice lists were presented at all testing SNR levels (i.e., -9, -7, -5 and -3 dB SNRs) in a random order to acquaint participants with the listening environment and the listening task before testing. The test sessions last about 2 hours for adult participants with rest included. The experimenter monitored participants' fatigue and attention, and offered breaks when necessary.

The 32 sentence lists were presented at the four SNRs with both test order and

the presentation SNR counterbalanced across subjects (see Supplemental Material 1). In this way, each sentence list was tested by 16 participants at each SNR level. Learning and/or fatigue effects were minimized as the testing order of all lists were counterbalanced across participants.

Results

Analysis of list equivalence

The performance scores of all lists were calculated as the proportion of correct recognition of the 50 keywords for a particular list at each SNR averaged across 16 participants (see Supplemental Material 2 for descriptive data). A two-way repeated measures analysis of variance (ANOVA) was performed on the scores with SNR level as a repeated-measure factor to investigate the equivalence of the 32 sentence lists. Significant main effects of list number ($F_{31,465} = 3.953$, $p < 0.001$) and SNR ($F_{3,45} = 5201.672$, $p < 0.001$) were observed. Post hoc Tukey Honestly Significant Difference (HSD) test revealed that the recognition scores in List K, List Q, List U and List Y differed significantly from other lists. The four non-equivalent lists were subsequently removed from further analysis.

Performance-SNR function

The SNR_{50} and slope of performance-SNR intelligibility functions for each remaining sentence list were obtained by fitting a logistic regression function using the following equation (Nissen et al., 2005):

$$p = \frac{\exp(a - b * SNR)}{1 + \exp(a - b * SNR)} \quad (1)$$

in which the dependent variable p (percentage correct) is the performance score, a is the regression intercept, and b is the regression slope. The slope (percentage per dB) of the

steepest portion of the curve for each list was calculated as *-b/4*. The speech recognition threshold (SNR_{50}) was calculated as the ratio of *a/b*.

For the 28 sentence lists, the mean and standard deviation of SNR_{50} thresholds were -5.9 ± 0.1 dB, and the slope of performance-SNR functions, 22.3 ± 1.5%/dB. Psychometric function curves of all lists are shown in Figure 1.

[Figure 1]

Variability of scores

All recognition scores in percentage were normalized for comparison using the rationalized arcsine unit (rau) transform (Studebaker, 1985; Sherbecoe & Studebaker, 2004). This transformation can achieve a linear and additive scale, and come up with rationalized arcsine values nearly equal to percentages over a broad central range of percentages. The scores of the 28 lists at four SNR conditions were transformed via the following equations:

$$\theta = arcsin\sqrt{\frac{X}{N+1}} + arcsin\sqrt{\frac{X+1}{N+1}} \quad (2)$$

$$R = \left(\frac{146}{\pi}\right)\theta - 23 \quad (3)$$

in which *X* represents the number of correct items, and *N* is the total number of tested items.

The standard deviations of normalized scores of each participant at each SNR were calculated. The grand SD was computed by averaging the squared standard deviations and taking the square root, resulting in a mean value of 9.8%. Given the slope of the P-SNR function (22.3%/dB), and the symmetrical distribution of scores around the 50% point, the corresponding inter-participant SD of SNR_{50} was 0.4 dB. By multiplying the grand SD by a coefficient of 2.77 (derived from √2×1.96 where 1.96 is

the appropriate *z* value for 95% confidence interval, and√2 indicates that the critical difference includes variation in two measurements), the 95% critical difference (CD) was subsequently calculated as 27.2% for score and 1.2 dB SNR for speech recognition threshold (SNR_{50}).

Experiment 2: List equivalency in pre-school children with normal hearing

Materials and methods

Materials

The 28 sentence lists observed to be equivalent in adults with normal hearing in Experiment 1 were administered to children with normal hearing. Since three SNR conditions were determined to be tested in pediatric participants, the total number of lists need to be adjusted to a multiple of three so that all lists can be counterbalanced across SNR conditions. Therefore, one list (List 6, with the greatest variance from other lists though insignificant) was removed, resulting in a total of 27 sentence lists tested in children with normal hearing. Each list consisted of 9 scoring sentences with 50 keywords, and a preliminary example sentence.

Participants

A total of 54 native Mandarin-speaking children (27 male, 27 female) aged 4–6 years were recruited from a kindergarten in Beijing to participate in Experiment 2. To balance the age of the pediatric participants, 18 children were recruited for each of three age group: 4 to 4.5 years old, 4.5 to 5 years old, and 5 to 6 years old. All children had no known history of otologic diseases, with pure tone thresholds for each audiometric frequency from 250 Hz to 8 kHz ≤25 dB HL and normal results of otoscopic examination and tympanometry. The recruitment of human subjects in the present study

was reviewed and approved by the Institutional Review Board of the Chinese PLA General Hospital, Beijing, China.

Experimental design and procedure

The pediatric participants attended two sessions because of the children's short attention spans and low cooperative motivation. In the first session, otoscopic examination, pure tone audiometry and tympanometry were administered to screen the pediatric participants, then the suitable children were familiarized with the speech test procedure. In the second session, the formal sentence recognition test was administered after at least two practice lists were completed. The children were instructed to listen carefully to each sentence after the beep and repeat as many words as possible. Performance was scored by the experimenter during the test in terms of number of keywords correctly repeated. The test session lasted an average of 1.5 hours with break periods included. The experimenter monitored children's fatigue and attention, and offered verbal encouragement, food reward, or breaks when necessary to encourage the children to complete all required testing.

The test was administered in a quiet room in a kindergarten. The ambient noise was ≤40 dB (A). The instruments (CD player, audiometer, and loudspeaker) used for the presentation of stimulus sentences and babble noise were the same as in Experiment 1. The sentences were presented at a fixed level of 65 dB SPL, slightly higher than the level in Experiment 1 given the higher level of ambient noise. The intensity of the babble noise was adjusted by the experimenter to achieve +1, -2, -5 dB SNRs. The SNR levels were pre-determined based on the results of a pilot test with six children. They were first tested at the highest SNR level used in the experiment with adults, i.e., -3 dB SNR, and scored around 35%. We then increased the SNR to -2 dB SNR, and their performance scores reached about 50%, suggesting that -2 dB SNR would produce

approximately 50% intelligibility for pediatric listeners with normal hearing. The loudspeaker was placed 1 m in front of the participants and approximately 30 cm from one corner of the room with its center at the same height as the ear of the child in a sitting position. The sound field was calibrated in the same way as in Experiment 1.

Considering children's limited attention span, only 15 of the 27 lists were tested for each child. The sentence lists were assigned in a balanced way to three age groups in the following way: first, children were divided into three age groups (aged 4-4;6, aged 4;6-5, and aged 5-6) with 18 participants in each group; then each list were distributed to be tested 10 times in each group (18 children × 15 lists = 27 lists × 10 times). The 15 sentence lists were presented at different SNRs, with the starting list varying across subjects (see Supplemental Material 3). In this way, all lists were optimally counterbalanced across SNRs resulting in each list tested by 10 different children in each of three SNR conditions. Learning and/or fatigue effects were minimized as the presentation order of the lists was also counterbalanced among the participants, with each list having equal probability of being presented first.

Results

Analysis of list equivalence

The performance scores were calculated as the proportion of correct recognition of the 50 keywords for each list at each SNR averaged across 10 participants (see Supplemental Material 4 for descriptive data). A two-way repeated measures ANOVA was performed using SNR level as a repeated-measure factor to investigate the equivalence of the 27 lists in pediatric participants. Significant main effects of list number ($F_{26,234} = 1.86$, $p = 0.009$) and SNR ($F_{2,18} = 6197.14$, $p < 0.001$) were observed for recognition scores. However, the follow-up Tukey's HSD post hoc test revealed no

344 significant differences among the 27 lists.

345 *Performance-SNR function*

346 The Performance-SNR psychometric functions for the 27 lists were obtained based on
347 the same mathematical procedures as in Experiment 1. Logistic curves were fitted for
348 the recognition scores of each list.

349 For the 27 lists, the mean and standard deviation of the slopes at the SNR_{50}
350 across the lists were 15.8 ± 1.1%/dB, and the SNR_{50}, -2.0 ± 0.2 dB. Psychometric
351 function curves of all lists are shown in Figure 2.

352 [Figure 2]

353 *Variability of scores*

354 All recognition scores of the 27 lists at three SNR conditions (n = 10) were normalized
355 by a rationalized arcsine transformation as in Experiment 1. The grand SD was
356 computed as 8.9% for score and 0.6 dB for SNR_{50}. The critical difference was
357 calculated in the same way as in Experiment 1, resulting in a 95% CD of 24.6% for
358 score and 1.6 dB for SNR_{50}.

359 *Age effect*

360 A simple linear regression was conducted to assess the effect of age on recognition
361 scores. For SNR = -5 dB, the regression equation was: score = 0.004 + 0.025 × age,
362 which was found to be statistically significant ($R^2 = 0.046$, $F_{1,268} = 12.944$, $p = .000$).
363 For SNR = -2 dB, the regression equation was: score = 0.112 + 0.08 × age, which was
364 found significant ($R^2 = 0.171$, $F_{1,268} = 55.349$, $p = .000$). For SNR = +1 dB, the
365 regression equation was: score = 0.557 + 0.063 × age, which was found significant (R^2

= 0.192, $F_{1,268}$ = 63.667, p = .000). Generally, children's age significantly predict their recognition scores in all three SNR conditions.

Discussion

This paper provides evidence on list equivalency and critical differences for the Mandarin BKB sentence in noise test for adults and young children with normal hearing. Sentences with paired babble from the BKB-like sentence corpus in Mandarin Chinese (HOPE corpus, Xi et al., 2012) were compiled into 32 lists with balanced distributions of syntactic difficulty and sentence type. For the 28 lists that were found to be equivalent in difficulty for adults, a critical difference of 1.2 dB for SNR_{50} was obtained at the speech presentation level of 60 dB SPL. Of the 28 lists, 27 were equivalent in difficulty for children, and a critical difference of 1.6 dB for SNR_{50} was obtained for the speech presentation level at 65 dB SPL. Using percent correct scores for the equivalent lists, the critical difference was 27.2% for adults and 24.6% for children. These values suggest that any greater difference can be interpreted as significant at the 95% confidence level.

By establishing list equivalency and critical differences, the present study provides a valuable resource for future applications. In clinical practice, speech perception tests usually include several equivalent lists for repeated measurements to control for potential learning effects. When different lists are administered at different intensity levels, one can estimate the speech intelligibility curve, provided that the lists are interchangeable. That is, any listener receiving a particular percentage correct score on one list under certain listening conditions should receive the same score (within certain limits) on any other list if all other variables remain constant.

In clinical trials, lists that are not equivalent in difficulty can be used to compare scores between two conditions, e.g., for within-subject assessment of the effects of

391 different forms of hearing prostheses (e.g., HAs or CIs) or training. The difference in
392 scores between two presentations must be large enough to be attributable to the
393 experimentally varied condition rather than to chance alone. The smallest difference
394 that must be exceeded for a difference between two scores is termed the critical
395 difference. Considering the experimental design of the current study, this variability was
396 quantified with inter- and intra-subject differences combined by calculating the standard
397 deviation of the resulting distributions of scores in a group of subjects (Dillon, 1982). In
398 this study, list scores for each listener were normalized before the calculation of critical
399 difference using the rationalized arcsine unit (rau) transform to achieve a linear and
400 additive scale. This transformation was proposed by Studerbaker (1985) for the data
401 analysis of speech recognition scores. The purposes of the rau transform are firstly so
402 that the random error component more closely approximates a Gaussian distribution
403 even when the score is close to 0% or 100%, secondly to minimize the correlation of
404 variability and performance, and thirdly so that the transformed scores approximate the
405 original proportion correct over most of the performance range. The mean values of
406 9.8% in adults and 8.9% in children represent the variability attributable to a
407 combination of the differences in true list difficulty and random measurement error
408 within a subject. The critical difference for two scores based on single lists to be
409 significantly different, with a 95% confidence interval, is thus 27.2% for adults and
410 24.6% for children. The critical difference is useful when evaluating the efficacy of a
411 treatment or rehabilitation. These values approximate the estimate for the Australian
412 version of BKB sentence lists (24.6%), which also contained 50 keywords in each list
413 (Keidser et al., 2002). In both cases, the standard deviation of repeated scores (e.g.,
414 9.8% for adults in this study) is slightly larger than the value of 7.1% that would be
415 expected from the binomial distribution for a true score 50% correct in a sample of 50

independent items (Hagerman, 1976). The larger value observed is undoubtedly because probabilities of correctly recognizing the keywords in a semantically rich sentence are not statistically independent of each other, so effectively there are fewer than 50 *independent* scored items per list.

Another significant issue that needs to be addressed in the design of standardized speech perception tests is sensitivity (Dillon & Ching, 1995). For a test to be sensitive to some specific change in listening conditions (e.g., SNR), the resulting change in scores must be larger than the variation expected on the basis of chance alone. Test materials with a steep slope of the P-SNR function are more sensitive than materials with shallower slopes because any small change in SNR is more likely to lead to a change larger than occur due to chance alone. Therefore, the steep slope of the P-SNR function is generally one of the focuses in the development of speech-in-noise tests. When testing the sentence lists in adults and children with normal hearing in the current study, the average slopes of P-SNR functions were 22.3 ± 1.5%/dB (28 lists) and 15.8 ± 1.1%/dB (27 lists), respectively. Compared with the slope of 10%/dB in the HINT (Nilsson et al., 1994) and 9%/dB in the MHINT (Wong et al., 2007), the current set of sentence-recognition-in-babble-noise lists has a steeper slope of the P-SNR function. This difference may be partly attributed to the use of steady-state competing noise in the HINT and MHINT, and partly to the relative low difficulty of the test material as the tests were designed for school-aged children. Test sensitivity and the slope of the psychometric function are maximized by making all items in the test equally difficult (Dillon, 1983). For our Mandarin BKB sentence test, item homogeneity was achieved through level adjustment of each keyword for scoring so that all words had the same SNR_{50} (Xi et al., 2012) following the ICRA recommendations for the construction of multilingual tests (Akeroyd et al., 2015). Such optimization based on word-specific

441 SNR_{50} has also been applied in the development of CMN matrix, resulting in a
442 relatively steep mean slope of 13.1%/dB (Hu et al., 2018). The results are consistent
443 with previous studies showing that the more homogeneous performance is on the
444 individual test items with respect to both location and slope, the steeper the slope of the
445 mean psychometric function (Wilson & Carter, 2001). This keyword-based
446 homogenization method contrasts with that used in the development of the MHINT test,
447 where sentence, rather than individual word, difficulty was equalized, suggesting that
448 the word-specific optimization of homogeneity can result in better sensitivity.
449 Accordingly, the current sentence test has higher sensitivity to differences between test
450 conditions and can thus evaluate different algorithms and settings of hearing prostheses
451 with respect to speech intelligibility in noisy environments with better accuracy and
452 efficiency (Wilson & Carter, 2001).

453 The strength of this study lies in establishing the list equivalency of 28 lists of
454 BKB like Mandarin sentences drawn from a corpus of sentences each with a paired
455 babble noise in which the keywords for scoring are homogeneity optimised. The 28 lists
456 can be used with adults, and 27 of them also with pre-school children. Although list
457 equivalency and critical differences have been established, this study does not provide
458 evidence for clinical applications with adults and children who are deaf or hard of
459 hearing. Future research needs to be carried out to investigate the clinical applicability
460 of the lists in evaluating performance with hearing prostheses, including hearing aids
461 and cochlear implants.

462 Unlike most existing speech perception tests in Mandarin Chinese that use
463 steady-state speech-shaped noise as competing sounds, the present Mandarin BKB lists
464 used 4-talker babble as competing sounds so that the test has greater face validity in
465 measuring an individual's ability to understand speech in everyday acoustic

environments for speech communication. The use of a sentence paired with a frozen babble allowed effective optimisation of homogeneity across keywords used for scoring (Xi et al, 2012). The 4-talker babble provides energetic as well as informational masking, while the sentences are natural utterances based on a corpus of utterances of children aged between 4 and 5 years (Liu et al., 2008). Accommodating the cognitive, motor, and attentional capabilities of young children, the sentence length was controlled to 6–8 characters, shorter than sentences in MHINT (10 characters; Wong et al., 2007) and CMN matrix (11 characters; Hu et al., 2018). A single list takes approximately 1.5 min to administer, which is appropriate for clinical use among children 4 years of age and older.

Although we have attempted to counterbalance lists and SNRs across participants in the study using a Latin square design (Supplementary Material 3), the order of SNRs was constant across participants during test administration. To reduce the potential effect of order on performance, each participant was tested at each SNR for five times such that testing at a certain SNR occurred sometimes at the beginning and other times towards the end of a test session. The potential effect of the order of SNRs will need to be quantified in future work.

The effect of age on performance observed in this study has been reported in other studies (e.g., the MHINT-C (Chen & Wong, 2020), the Cantonese Hearing in Noise Test for Children (CHINT-C; Wong, Chen & Leung, 2019), the Canadian French HINT for children (Vaillancourt et al., 2008). Future research will attempt to replicate the present findings on list equivalency and examine test-retest reliability in children across a wide age range.

Conclusions

The Mandarin BKB sentences in noise test offers an opportunity for clinicians and researchers to assess speech understanding in adults and pre-school children in an efficient manner. For clinical comparisons of performance with different presentation conditions or settings, 28 equivalent lists are available for adults, and 27 lists for preschool children. These lists can be scored using keyword scoring. Furthermore, the 95% critical difference values can be used for total percentage correct or SNR for 50% performance to assess the effectiveness of listening in real-life environments. Future work on this research tool should examine clinical applicability for school-aged children and children who are deaf and hard of hearing.

Author Note

Portions of this article were previously described in Chinese in the following article and permission for reuse has been obtained from the publisher:

Xi, X., Chen, A. T., Li, J. N., Ji, F., Hong, M. D., Yang, S. M., Han, D.Y. (2009). Caozayu zaosheng xia Putonghua ertong yuju ceting biao de biaozhunhua [Standardized Mandarin Sentence Perception in Babble Noise Test Materials for Children]. *Journal of Audiology and Speech Pathology*, *17*(4), 318-322.

Acknowledgments

The authors acknowledge Mr. Yang Zhao's efforts in adults' data collection and the support of the NIHR Manchester Biomedical Research Centre.

Funding

This research is sponsored by National Key R&D Program of China (Grant No. 2020YFC2004005, 2007BAI18B12, 2008BAI50B01), and supported by Science Foundation of Beijing Language and Culture University (the Fundamental Research Funds for the Central Universities, 21PT01).

ORCID ID

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

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

Disclosure Statement

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

Data Availability Statement

The datasets generated and analyzed during the current study, as well as the recorded speech materials and instructions for test administration, are available from the first author, Dr. Xin Xi via email xixin_plagh@yeah.net upon reasonable request.

References:

Akeroyd, M. A., Arlinger, S., Bentler, R. A., Boothroyd, A., Dillier, N., Dreschler, W. A., Gagné, J. P., Lutman, M., Wouters, J., Wong, L., Kollmeier, B., & International Collegium of Rehabilitative Audiology Working Group on Multilingual Speech Tests. (2015). International Collegium of Rehabilitative Audiology (ICRA) recommendations for the construction of multilingual speech tests. ICRA Working Group on Multilingual Speech Tests. *International Journal of Audiology*, *54 Suppl 2*, 17–22.

American Speech-Language-Hearing Association. (1991). Sound field measurement tutorial. *Asha 33*(Suppl. 3), 25–37.

Bench, J. & Bamford. (1979). *Speech-Hearing Tests and the Spoken Language of Hearing-Impaired Children*. London, Academic Press.

Byrne, D., Dillon, H., Tran, K.V., Arlinger, S., Wilbraham, K., Cox, R.M., Hagerman, B., Hétu, R., Kei, J., Lui, C.P., Kiessling, J., Kotby, M.N., Nasser, N.H., Kholy, W.A., Nakanishi, Y., Oyer, H.J., Powell, R., Stephens, D., Meredith, R., Sirimanna, T., Tavartkiladze, G., Frolenkov, G.I., Westerman, S., & Ludvigsen, C. (1994). An international comparison of long-term average speech spectra. *Journal of the Acoustical Society of America*, *96*, 2108-2120.

Chen, Y., & Wong, L. L. N. (2020). Development of the Mandarin hearing in noise test for children. *International Journal of Audiology*, *59*(9), 707–712.

China State Language Commission. (2021). *The Report on the Development of Chinese Language*. Beijing, The Commercial Press.

Ching, T. Y. C., Dillon, H., Leigh, G., & Cupples, L. (2018). Learning from the Longitudinal Outcomes of Children with Hearing Impairment (LOCHI) study: summary of 5-year findings and implications. *International Journal of Audiology*, *57*(sup2), S105-S111.

Dillon, H. (1982). A quantitative examination of the sources of speech discrimination test score variability. *Ear and Hearing* (3), 51–58.

Dillon H. (1983). The effect of test difficulty on the sensitivity of speech discrimination tests. *The Journal of the Acoustical Society of America*, *73*(1), 336–344.

Dillon, H. & Ching, T. (1995). What makes a good speech test? in G. Plant & K. E. Spens (Ed). *Profound Deafness and Speech Communication*. London, Whurr Publishers: 305–344.

Etymotic Research. (2005). BKB-SIN Speech-in-Noise Test (Compact Disk), Elk Grove Village, IL.

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*(6), EL267–EL273.

Hagerman, B. (1976). Reliability in the determination of speech discrimination. *Scandinavian Audiology 5*: 219-228.

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*(11), 838–850.

Keidser, G., Ching, T.Y., Dillon, H., Agung, K., Brew, C., Brewer, S., Fisher, M.J., Foster, L., Grant, F., & Storey, L. (2002). The National Acoustic Laboratories (NAL) CDs of speech and noise for evaluation: normative data and potential applications. *Australian and New Zealand Journal of Audiology*, *24*, 16–35.

Killion, M. C., Niquette, P. A., Gudmundsen, G. I., Revit, L. J., & Banerjee, S. (2004). Development of a quick speech-in-noise test for measuring signal-to-noise ratio loss in normal-hearing and hearing-impaired listeners. *The Journal of the Acoustical Society of America*, *116*(4 Pt 1), 2395–2405.

Leibold, L. J., & Buss, E. (2019). Masked Speech Recognition in School-Age Children. *Front. Psychol.* 10, 1981.

Li, J. N., Chen, S., Zhai, L., Han, D. Y., Eshraghi, A. A., Feng, Y., Yang, S. M., & Liu, X. Z. (2017). The advances in hearing rehabilitation and cochlear implants in China. *Ear and Hearing*, *38*(6), 647–652.

Litovsky, R. Y., Johnstone, P. M., & Godar, S. P. (2006). Benefits of bilateral cochlear implants and/or hearing aids in children. International journal of audiology, 45 Suppl 1(Suppl 1), S78–S91.

Liu, S., Han, D.M., Zhang, N., Wu, X., Sheng, Y. L., Mo, L.Y., Yang, X.L., Kong, Y. (2008). Development of a database of everyday speech material for preschool Mandarin children (in Chinese). *Journal of Audiology and Speech Pathology* (China), *16*, 121–124

Moore, B.C. (2003). Speech processing for the hearing-impaired: successes, failures, and implications for speech mechanisms. *Speech Communication*, *41*, 81-91.

596 Minimum-Speech-Test-Battery. (2011) The new minimum speech test battery.
597 http://auditorypotential.com/MSTB.html.
598 Newman R.S., Morini G, and Ahsan F. (2015) Linguistically-based informational
599 masking in preschool children. *The Journal of the Acoustical Society of*
600 *America 138(1), EL93-E198*.
601 Nilsson, M., Soli, S. D. & Sullivan, J. (1994). Development of the hearing in noise test
602 for the measurement of speech reception thresholds in quiet and in noise.
603 *Journal of the Acoustical Society of America 95*, 1085–1099.
604 Nissen, S. L., Harris, R. W., Jennings, L. J., Eggett, D. L., & Buck, H. (2005).
605 Psychometrically equivalent Mandarin bisyllabic speech discrimination
606 materials spoken by male and female talkers. *International Journal of*
607 *Audiology*, *44*(7), 379–390.
608 Shaughnessy, J. J., Zechmeister, E. B., & Zechmeister, J. S. (2006). *Research methods*
609 *in psychology*. New York, McGraw Hill.
610 Sherbecoe, Robert L. & Studebaker, Gerald A. (2004). Supplementary formulas and
611 tables for calculating and interconverting speech recognition scores in
612 transformed arcsine units. *International Journal of Audiology, 43*(8), 442 - 448
613 Studebaker, G. A. (1985). A "rationalized" arcsine transform. *Journal of Speech and*
614 *Hearing Research*, *28*(3), 455–462.
615 Vaillancourt, V., Laroche, C., Giguère, C., & Soli, S. D. (2008). Establishment of age-
616 specific normative data for the Canadian French version of the hearing in noise
617 test for children. *Ear and Hearing*, *29*(3), 453–466.
618 Wilson, R.H. (2004). Adding speech-in-noise testing to your clinical protocol: Why and
619 how. *The Hearing Journal*, 57, 10.
620 Wilson, R. H., & Carter, A. S. (2001). Relation between slopes of word recognition
621 psychometric functions and homogeneity of the stimulus materials. *Journal of*
622 *the American Academy of Audiology*, *12*(1), 7–14.
623 Wong, L.L, Soli, S.D., Liu, S., Han, N. & Huang, M.W. (2007). Development of the
624 Mandarin hearing in noise test (MHINT). *Ear and Hearing*, *28*, 70–74.
625 Wong, L. L. N., Chen, Y., & Leung, K. P. (2019). The Cantonese hearing in noise test
626 for children. *Trends in Hearing*, *23*, 1-9.
627 Xi, X., Ching, T. Y., Ji, F., Zhao, Y., Li, J. N., Seymour, J., Hong, M. D., Chen, A. T.,
628 & Dillon, H. (2012). Development of a corpus of Mandarin sentences in babble

with homogeneity optimized via psychometric evaluation. *International Journal of Audiology*, *51*(5), 399–404.

Zheng, Y., Soli, D.S, Wang, K., Meng, J., Meng, Z.-L. (2009). Development of the Mandarin Pediatric Speech Intelligibility (MPSI) test. *International Journal of Audiology*, *48*, 718–728.

Zhu, M.S. & Miao, X.C. (eds.) (1990). *Psycholinguistics*. Shanghai, East China Normal University Press.

Figure 1. The cluster of Performance-SNR functions for the 28 sentence lists in babble noise in adults with normal hearing.

Figure 2. The cluster of Performance-SNR functions for the 27 sentence lists in babble noise in children with normal hearing.

Figure 3. The scatter diagram of children's performance scores at 3 SNRs with linear regressions of age and performance at 3 SNRs.

Supplemental Material 1. The SNR levels at which each of the 32 lists were tested for each subject, with the number of measurements at -3, -5, -7, -9, respectively, and the total number of measurements at four SNR levels for each list.

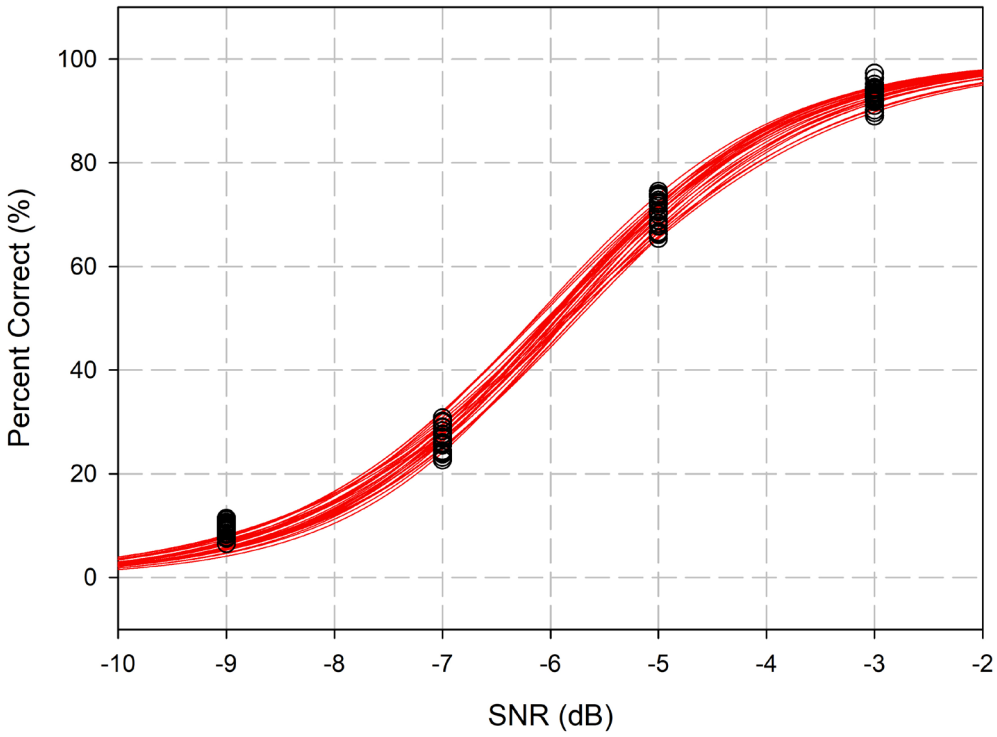
Supplemental Material 2. The mean and SD of the 32 lists tested in NH adults at four SNRs. Lists that were significantly different from others according to ANOVA results were italicized.

Supplemental Material 3. The SNR levels at which each of the 27 lists were tested for each subject, with the number of measurements at +1, -2, -5, respectively, and the total number of measurements at three SNR levels for each list.

Supplemental Material 4. The mean and SD of the 27 lists tested in NH children at three SNRs.

660 Figure 1.

661



662

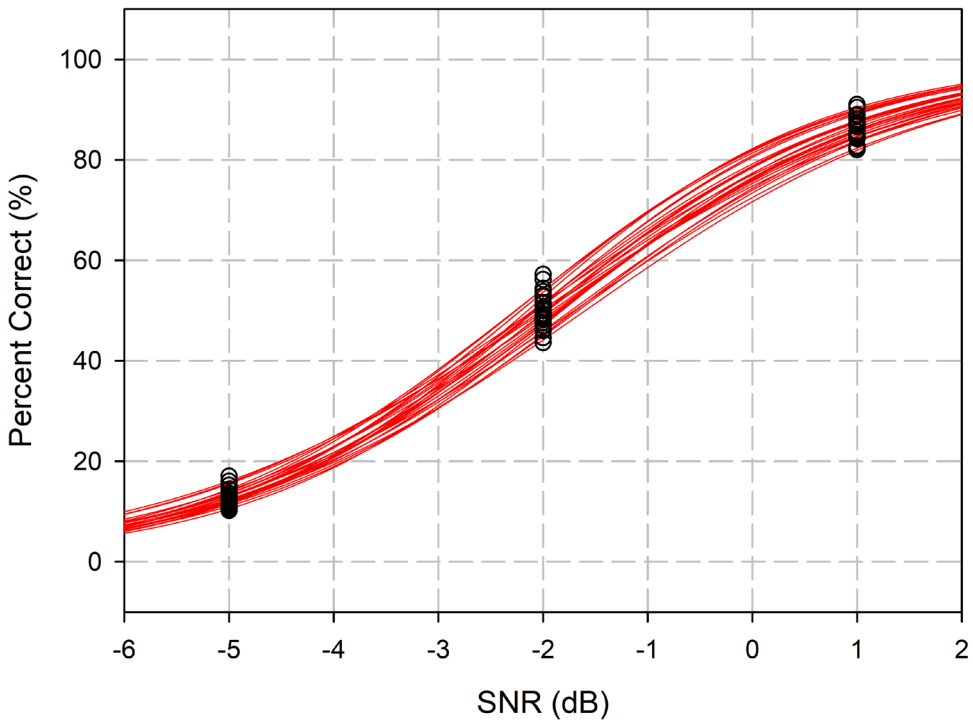
663

664

665

666 Figure 2.

667

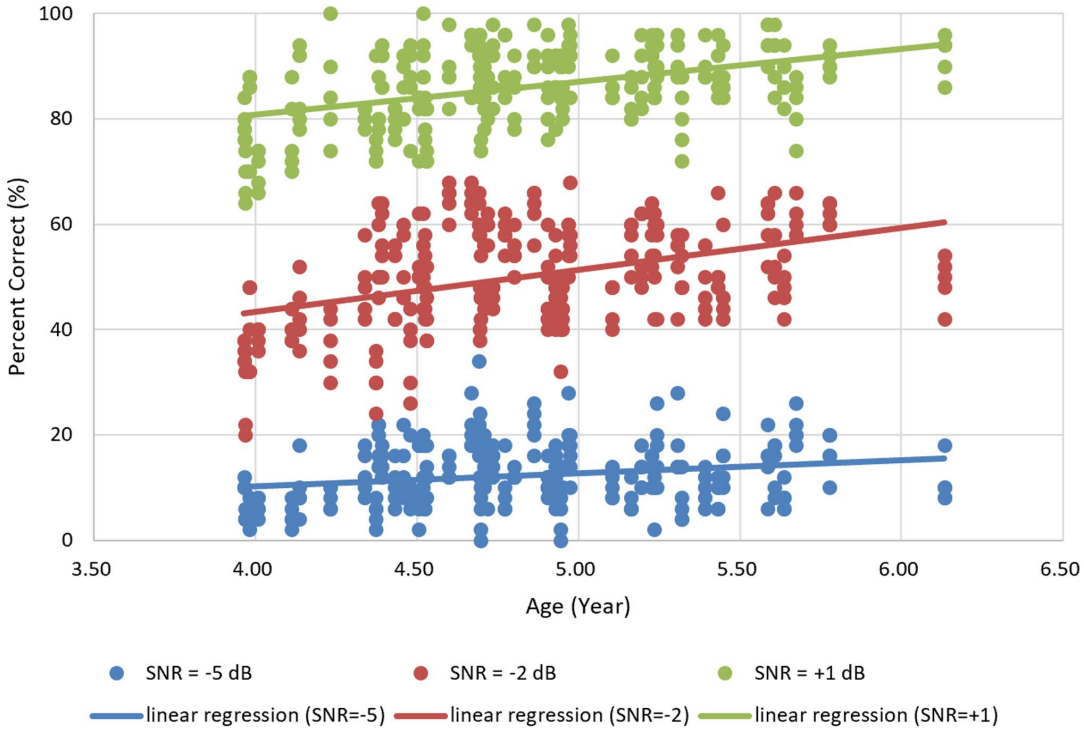


668

669

670 Figure 3.

671



672