Jurnal Sains Kesihatan Malaysia **19** (1) 2021: 7 - 19
DOI : http://dx.doi.org/10.17576/JSKM-2021-1901-02

Kertas Asli/*Original Article*s

Development of Digitized Mandarin Paediatric Speech Perception Test Materials for Malaysian Children

(Pembangunan Bahan Ujian Persepsi Pertuturan Pediatrik Mandarin Digital untuk Kanak-kanak di Malaysia)

ELIZABETH SHU JUN LIM, FOONG YEN CHONG*, TIAN KAR QUAR, HUI WOAN LIM, CILA UMAT, LEE LEE SOH & YEE VEN LEU

ABSTRACT

Digitized Mandarin paediatric speech perception tests are limited in Malaysia for measuring outcomes among children fitted with amplification devices. Mandarin speech perception tests from other countries may not be suitable to be used in Malaysia due to regional vocabulary differences. In this study, we aim to develop digitized test materials to test Mandarin-speaking preschool children in Malaysia. This is a two-phase cross-sectional study. Phase I is where Mandarin words (n=113) with test item pictures were collated and developed and then tested on 40 Mandarin-speaking children (aged 2;0 to 5;11 years old) with normal hearing. Phase II is where the Mandarin words were digitally recorded and then followed by acoustic analysis, sound quality evaluation, and validation by 20 Mandarin-speaking young adults. In Phase I, a total of 80 of 113 words with high familiarity among children were selected. In Phase II, a total of 192 recorded tokens were selected as finalized test items and all pictures corresponding to these items were validated as suitable representation of the test items. In conclusion, a Mandarin speech perception test which consists of digitally recorded stimuli and pictures has been developed for Mandarin-speaking pre-school children in Malaysia. Objective and subjective evaluations are important to select the best digitized test items and pictures for developing a speech perception test. Future research includes collecting normative data for the test and evaluating the test application in local audiology clinics.

Keywords: Mandarin; paediatric; speech perception test; digitized stimuli; outcome measures

ABSTRAK

Ujian persepsi pertuturan pediatrik dalam bahasa Mandarin yang berbentuk digital adalah amat terhad di Malaysia untuk menilai kanak-kanak yang memakai alat bantu pendengaran. Ujian persepsi pertuturan bahasa Mandarin dari negara lain adalah kurang sesuai digunakan di Malaysia disebabkan perbezaan perbendaharaan kata serantau. Kajian ini bertujuan untuk membina bahan ujian digital untuk menilai kanak-kanak prasekolah yang menutur bahasa Mandarin di Malaysia. Ini merupakan kajian dua fasa yang berbentuk keratan rentas. Fasa I melibatkan pengumpulan perkataan Mandarin (n=113) dengan gambar item ujian dan seterusnya dinilai ke atas 40 kanak-kanak penutur bahasa Mandarin (berumur 2;0 hingga 5;11 tahun) yang berpendengaran normal. Fasa II merangkumi rakaman digital perkataan dan diikuti dengan analisis akustik, penilaian kualiti bunyi, serta validasi oleh 20 orang dewasa muda penutur bahasa Mandarin. Dalam Fasa I, sejumlah 80 daripada 113 perkataan yang mudah dikenali oleh kanak-kanak telah dipilih. Dalam Fasa II, sejumlah 192 token rakaman telah dipilih sebagai item ujian dan setiap gambar yang mewakili item ujian dinilai sebagai lambang yang sesuai. Kesimpulannya, satu ujian persepsi pertuturan bahasa Mandarin yang merangkumi stimulus rakaman digital dan gambar telah dibina bagi kanak-kanak prasekolah yang menutur bahasa Mandarin di Malaysia. Penilaian objektif dan subjektif adalah penting untuk memilih item ujian digital dan gambar yang bagus bagi membentuk suatu ujian persepsi pertuturan. Cadangan kajian masa depan termasuk pengumpulan data normatif untuk ujian tersebut dan penilaian aplikasi ujian tersebut di klinik audiologi tempatan.

Kata Kunci: Bahasa Mandarin; pediatrik; ujian persepsi pertuturan; rangsangan digital; ukuran hasil intervensi

INTRODUCTION

Paediatric speech perception tests are vital objective assessment tools for evaluating speech comprehension and auditory skills among children (Mendel 2009; Millett 2012). These tests are also used in aiding early treatment or intervention of hearing loss among paediatric population (Zheng et al. 2009) such as providing information to determine the effectiveness of amplification devices fitted on children (Eisenberg et al. 2005). This helps in planning appropriate habilitation strategies which are customizable for each individual child. Examples of materials used in paediatric speech perception tests are syllables, words and sentences. These speech materials are complex signals and are appropriate to test children's speech perception ability in daily life (Mendel & Danhauer 1997; Ramkissoon 2001). Speech perception ability is important to ensure normal speech, language, cognitive, psychology and psychosocial development among children (Bess et al. 1998; Nelson et al. 2008; Theunissen et al. 2015).

In evaluating the children's speech perception ability, cultural and language background factors need to be considered as they may affect the children's performance in the test and may not reflect their true performance (Eisenberg et al 2005; Marinova-Todd et al. 2011; Mendel 2008). Many speech materials have been developed and validated for use in children who speak English as their first or dominant language. In other countries such as China where Mandarin is the official language, speech materials too have been developed and standardized in the said language to assess speech perception ability among Chinese children. One example is the Mandarin Early Speech Perception (MESP) test that is developed for young children in Mainland China (Zheng et al. 2009). The MESP test contains six test domains in a hierarchical manner (i.e., speech sound detection, speech pattern perception, spondee perception, consonant perception, vowel perception and tone perception). The test has been validated and is reliable to test children aged between two to six years old in Mainland China. Nonetheless, the MESP test may not be appropriate for evaluating other Chinese children who are born and raised in other countries due to the differences in culture and variation in regional vocabulary. For example, vocabulary differences exist between the standard Mandarin language (*Putonghua*) in China and Mandarin language (*Huayu*) in Malaysia (Leitner et al. 2016). Test items in MESP such as 自行车 /zì xíng chē/ (bicycle), 燕 /yàn/ (swallow) and 袜袜 /wà wa/ (sock) are more commonly known as 脚车 /jiǎo chē/ (bicycle),吞 /tūn/ (swallow) and 袜子 /wà zi/ (sock) in Malaysia. As such, there is a demand for Mandarin speech perception tests among the Mandarin-speaking young children in Malaysia and perhaps other neighbouring countries in the South East Asia region such as Singapore, Indonesia, Brunei and Thailand where Mandarin is also spoken. In Malaysia, Mandarin is one of the languages commonly spoken by the Chinese ethnic group, the second largest ethnic group which comprised of 23.0% of the Malaysia's population (Department of Statistics Malaysia 2018). The Malaysian Chinese ethnic children normally use Mandarin (and Chinese dialects) at home. For these children, their acquisition and proficiency of languages other than their mother tongue may be better achieved when they begin their formal education at schools (Hendrayani 2015).

There are limited tests in Chinese which cater to the local Malaysian Chinese children. To date, existing standardized Mandarin speech perception tests developed in Malaysia include the Syllabic Pattern Perception Test (SPPT) with Tone Perception test (TPT) (Umat et al. 2010), Mandarin Closed-Set Sentence Test (Ang 2010) and the Mandarin Fricative-Affricate Nonsense Word Test (Chong et al. 2018; Chong et al. 2020). The SPPT aims to assess speech pattern perception and consisted of monosyllabic, disyllabic and trisyllabic test items. The TPT aims to assess Mandarin tone perception and consists of several minimal pairs of monosyllabic words with tonal contrasts. These tests are suitable for children aged three to six years old. The Mandarin Closed-Set Sentence Test (Ang 2010) aims to assess closed-set sentences perception and is suitable for evaluating children aged four to six years old, however, it is not suitable for younger children. The Mandarin Fricative-Affricate Nonsense Word Test (Chong et al. 2018; Chong et al. 2020) aims to assess the perception of Mandarin fricatives and affricates among Mandarin-speaking adults in Malaysia and there is no normative data for children. Therefore, the number of Mandarin tests available in Malaysia for testing young children is considered limited as compared to countries where Mandarin is widely spoken (e.g., China and Taiwan).

In addition, these locally developed paediatric speech perception tests have several limitations when compared to the Mandarin Early Speech Perception (MESP) developed in Mainland China mentioned earlier. First, each of the SPPT and TPT tests (Umat et al. 2010) only contains one test domain (i.e., speech pattern perception and tone perception, respectively). In contrasts, the MESP test contains six test domains arranged in a hierarchical manner. These test domains could assess different auditory perception aspects and provide a more comprehensive representation of children's speech perception skills. Second, the MESP test (Zheng et al. 2009) has more test items as compared to the TPT test (Umat et al. 2010). In MESP test, there are 24 minimal pairs (four minimal pairs for each of the six Mandarin tonal contrasts) for assessing tone perception whereas in TPT, there are only six minimal pairs (one minimal

pair for each tonal contrast). The higher number of test items may improve the reliability of a test due to a decreased chance score. With regards to the scarcity of Mandarin speech perception tests and the limitations in existing tests, the development of a new Mandarin speech perception test is indeed necessary.

The aim of this study was to develop materials which included test items, pictures and digitally recorded audio stimuli for a Malaysian Mandarin paediatric speech perception test. Two research questions are postulated:

1. What are the most familiar and suitable Mandarin words with pictures for testing speech perception among three- to six-year-old Mandarin-speaking children in Malaysia?
2. Which are the most suitable recorded test stimuli and pictures for developing a Mandarin paediatric speech perception test in Malaysia?

Specifically, the current study was conducted in two phases. Phase I focused on gathering vocabulary from various resources and selecting high familiarity words among Malaysian Mandarin-speaking preschool children as test items. Phase II focused on (i) developing digitized recorded word tokens for the vocabulary selected in Phase I, (ii) examining suitability of pictures as representation of the word tokens, and (iii) validating the recorded word tokens as finalized test items.

MATERIALS AND METHODS

This two-phase cross-sectional study received ethics approval (UKM1.21.3/244/NN-2017-043 and UKM PPI/111/8/JEP-2017-692) from the Universiti Kebangsaan Malaysia (UKM) Research Ethics Committee.

PHASE I: SELECTION OF TEST ITEMS

PARTICIPANTS

Forty Chinese descent Malaysian children (17 males, 23 females) aged two to five years old participated in this study. Participants were recruited by purposive sampling at nurseries and kindergartens located around Klang Valley. The inclusion criteria of participants were (1) Malaysian Chinese children (2) aged two to five years old and (3) used Mandarin mainly for daily communication. The exclusion criteria were children who (1) failed hearing screening, (2) had speech and physical development delay as reported by parents, (3) had syndromic disorders, autism, or cerebral palsy as reported by parents, and (4) were not of Malaysian citizenship.

MATERIALS

A list of Mandarin words were collated (n=113) from various sources namely: (1) Mandarin First Word Checklist (Chok 2001), (2) Mandarin Early Speech Perception Test (Zheng et al. 2009), (3) Multilingual English-Mandarin-Malay Phonological Test Scoring Form-Mandarin (Lim 2010), (4) Syllabic Pattern Perception Test and Tone Perception Test (Umat et al. 2010) and, (5) Mandarin Closed-Set Sentence Test (Ang 2010). The list consisted 73 monosyllabic words, 36 disyllabic trochees or spondees, and four trisyllabic words. These words were common nouns or verbs which can be represented in coloured pictorial form. Pictures representing the words were hand-drawn by the authors and reviewed by an expert panel consisting of two audiologists and one speech language pathologist.

The collated words were distributed into five test categories based on the criteria stated in Table I. The

TABLE 1. Criteria of word selection for test categories in Malaysian Mandarin Paediatric Speech Perception Test (MyMaPS)

Category	Structure
1 (SpPPT)	Words which are grouped according to the number of syllable (monosyllables, disyllabic trochees, disyllabic spondees and trisyllables)
2 (SPT)	Disyllabic spondees which are divided into three subcategories which are grouped according to the syllables tone
3 (VPT)	Monosyllables which are divided into four subcategories with similar consonant and tone but different vowels
4 (CPT)	Monosyllables which are divided into four subcategories with similar vowel and tone but different initial consonants
5 (TPT)	Minimally contrastive word pairs of six Mandarin tonal contrasts (Tone 1-Tone 2, Tone 1-Tone 3, Tone 1- Tone 4, Tone 2- Tone 3, Tone 2- Tone 4, and Tone 3- Tone 4 contrasts).

Note: SpPPT = Speech pattern perception test, SPT = Spondee perception test, VPT = Vowel perception test, CPT = Consonant perception test, TPT = Tone perception test.

five test categories are (1) Category 1 - speech pattern perception test (SpPPT), (2) Category 2 - spondee perception test (SPT), (3) Category 3 - vowel perception test (VPT), (4) Category 4 - consonant perception test (CPT), and (5) Category 5 - tone perception test (TPT). These categories are structured similarly to the MESP test categories (Zheng et al. 2009).

PROCEDURES

The field test was aimed at assessing word familiarity among the participants based on the picture-pointing task. Hearing screening was conducted to ensure that the children had normal hearing. The test was administered by the tester using live voice at normal conversational level without visual cues. When the child was viewing sets of five or 10 pictures, the words were presented randomly in order to avoid participants predicting the sequence of words presented. Each word was presented only once. The child was required to point to a picture corresponding to the word they heard. Reinforcements and breaks were given throughout the testing to maintain the child's attention. The identification score (the number of times a token was correctly identified per total number of participants X 100%) of each word was calculated. Words with the highest scores within each test category were considered as having high familiarity and were selected as materials for Phase II.

PHASE II: DEVELOPMENT OF DIGITIZED WORD TOKENS AND SUITABILITY OF PICTURES

PARTICIPANTS

Two undergraduate students (one male and one female) participated in the recording of word tokens. They did not have hoarse or breathy voice during the recording. Three professionals (two audiologists and one speech-language pathologist) participated in the evaluation of sound quality of the recorded words. They had prior phonetic training with a minimum five years of working experience. Another 20 undergraduate students (mean age=22.6 years old; age range=19-25 years old) participated in the validation of the recorded word tokens and corresponding pictures. All participants were native speakers of Mandarin and had normal hearing.

MATERIALS

The high familiarity words with pictures from Phase I were used.

PROCEDURES

The recording of word tokens was conducted in a semi-professional recording studio. During the recording session, each word selected in Phase I was uttered with a carrier phrase "I will say…" in Mandarin by each of the two talkers. Each word was uttered three times by the talkers. The talkers were instructed to use their usual conversational speech rate and voice during recording. The utterances were saved as Waveform Audio Format (WAV) at 32-bit and 44100 Hz in computer for offline processing. Following that, all recorded word tokens were extracted from their carrier phrases by one of the authors. A second author extracted 20% (96 tokens) of the recorded word tokens. Intraclass Correlation Coefficient (ICC) was performed to check the reliability between the token duration extracted by the two testers.

Subsequently, the tokens underwent acoustic analysis and visual inspections of their waveforms and spectrograms. This was to ensure that they are free from idiosyncrasies (Cheesman & Jamieson 1996) such as atypical pitch contour, different sound intensity, or irregularity in pronunciation (Chong et al. 2018). The acoustic analysis and visual inspection were also carried out by another author on 20% of the tokens that were randomly selected. The analysis of both of the authors were compared and tokens that did not meet the pitch contour criteria of Mandarin tones (Liu et al. 2007) were excluded. Following that, the selected tokens underwent sound quality ratings by two professionals. Tokens that were rated as good quality by two professionals were accepted whereas tokens that were rated as poor quality were excluded.

Suitability of pictures as representation of the selected word tokens was examined via a picture-naming test. Participants were presented with pictures corresponding to the recorded word tokens and they were required to name the pictures, one at a time. The number of times pictures being named correctly were calculated and the average score was utilized as the benchmark value to determine the suitability of each picture. Pictures with scores that were lower than the average were modified based on participants' feedback. This is to ensure that they are more representative of the recorded word tokens.

Validation of the recorded word tokens were examined through speech identification in a sound-treated room. The recorded word tokens were presented to the participants monaurally via an insert earphone at 60 dB HL to ensure adequate loudness and listening comfort. The participants were required to identify pictures which corresponded to the word tokens heard. The correct identification score (the number of times a token was correctly identified per total number of participants X 100%) was then calculated. Recorded word tokens that achieved the highest correct

TABLE 2. Words selected as test items for Category 1 to 4 and the respective identification score (%)

Test Categories	High Familiarity Words with Correct Identification Score (%)											
Category 1	Monosyllables			Disyllabic Trochees			Disyllabic Spondees			Trisyllables		
Score (%)	97.5	97.5	95.0	100.0	100.0	97.5	100.0	100.0	82.5	100.0	97.5	87.5
Chinese character	猫	狗	饭	椅子	鼻子	裤子	书包	飞机	花生	电视机	垃圾桶	红毛丹
Pinyin	/māo/	/gǒu/	/fàn/	/yǐ zi/	/bí zi/	/kù zi/	/shū bāo/	/fēi jī/	/huā shēng/	/diàn shì jī/	/lā jī tǒng/	/hóng máo dān/
English words	Cat	Dog	Rice	Chair	Nose	Trouser	Bag	Airplane	Peanut	Television	Dustbin	Rambutan
Category 2	Tone 1			Tone 2					Tone 4			
Score (%)	95.0	92.5	85.0	-	97.5	92.5	85.0	-	100.0	100.0	95.0	92.5
Chinese character	医生	香蕉	青蛙	西瓜	葡萄	蝴蝶	毛虫	榴莲	睡觉	电话	大象	月亮
Pinyin	/yī sheng/	/xiāng jiāo/	/qīng wā/	/xī guā/	/pú táo/	/hú dié/	/máo chóng/	/líu lián/	/shuì jiào/	/diàn huà/	/dà xiàng/	/yuè liàng/
English words	Doctor	Banana	Frog	Watermelon	Grape	Butterfly	Caterpillar	Durian	Sleeping	Telephone	Elephant	Moon
Category 3	Tone 1			Tone 2			Tone 3			Tone 4		
Score (%)	97.5	95.0	87.5	100.0	85.0	57.5	100.0	100.0	65.0	95.0	80.0	77.5
Chinese character	车	吃	叉	鱼	牙	椰	手	水	鼠	面	木	帽
Pinyin	/chē/	/chī/	/chā/	/yú/	/yá/	/yé/	/shǒu/	/shuǐ/	/shǔ/	/miàn/	/mù/	/mào/
English words	Car	Eat	Fork	Fish	Teeth	Coconut	Hand	Water	Mouse	Noodle	Wood	Hat
Category 4	Tone 1			Tone 2			Tone 3			Tone 4		
Score (%)	87.5	80.0	70.0	95.0	65.0	60.0	72.5	65.0	60.0	95.0	95.0	65.0
Chinese character	叉	鸭	八	羊	狼	糖	土	鼠	鼓	蛋	饭	站
Pinyin	/chā/	/yā/	/bā/	/yáng/	/láng/	/táng/	/tǔ/	/shǔ/	/gǔ/	/dàn/	/fàn/	/zhàn/
English words	Fork	Duck	Eight	Sheep	Wolf	Sugar	Soil	Mouse	Drum	Egg	Rice	Stand

identification score were selected as the finalized test items for the speech perception test.

RESULTS

PHASE I: SELECTION OF TEST ITEMS

The 113 words collated from various sources consisted of 73 monosyllabic words, 36 disyllabic words, and four trisyllabic words. The identification scores for the monosyllabic words were 80% or more for 38 words, between 60%-79% for 23 words, 59% or less for the remaining 12 words. The identification scores for the disyllabic words were 82.5% or more for 28 words, between 70%-77.5% for six words while another two words had the lowest percentage of 60%. All trisyllabic words were scored between 85%-100%.

The Malaysian Mandarin Paediatric Speech Perception Test (MyMaPS) comprises five test categories as in Table 1. The words that were selected to form each test category are tabulated in Table 2 and Table 3. For Category 1 (Speech Pattern Perception Test – SpPPT), a total of 24 words were tested and 12 words were selected. For Category 2 (Spondee Perception Test – SPT), a total of 19 words were tested and only 10 words were selected. Due to low identification scores (30-55%) for other words that were tested, the words **西瓜** /xī guā/ (watermelon) and **榴莲** /

TABLE 3. Minimal pairs selected as test items for Category 5 (Tone Perception Test) and the respective identification score (%)

Subcategories	Minimal Pairs with Correct Identification Score (%)							
	Pair #1		Pair #2		Pair #3		Pair #4	
Tone 1 - Tone 2								
Score (%)	88.8		86.3		82.5		75.0	
Chinese character	窗	床	飞	肥	鸭	牙	汤	糖
Pinyin	/chuāng/	/chuáng/	/fēi/	/féi/	/yā/	/yá/	/tāng/	/táng/
English words	Window	Bed	Fly	Fat	Duck	Teeth	Soup	Sugar
Tone 1 - Tone 3								
Score (%)	83.8		81.3		75.0		68.8	
Chinese character	车	尺	冰	饼	烟	眼	妈	马
Pinyin	/chē/	/chě/	/bīng/	/bǐng/	/yān/	/yǎn/	/mā/	/mǎ/
English words	Car	Ruler	Ice	Biscuit	Smoke	Eye	Mother	Horse
Tone 1 - Tone 4								
Score (%)	92.5		87.5		78.8		66.3	
Chinese character	眼睛	眼镜	猫	帽	哭	裤	书	树
Pinyin	/yǎn jīng/	/yǎn jìng/	/māo/	/mào/	/kū/	/kù/	/shū/	/shù/
English words	Eye	Spectacle	Cat	Hat	Cry	Trouser	Book	Tree
Tone 2 - Tone 3								
Score (%)	88.8		88.8		67.5		56.3	
Chinese character	鱼	雨	鞋	写	鼻	笔	图	土
Pinyin	/yú/	/yǔ/	/xié/	/xiě/	/bí/	/bǐ/	/tú/	/tǔ/
English words	Fish	Rain	Shoes	Write	Nose	Pencil	Picture	Soil
Tone 2 - Tone 4								
Score (%)	75.0		72.5		70.0		39.4	
Chinese character	毛	帽	鞋	蟹	河	喝	雷	泪
Pinyin	/máo/	/mào/	/xié/	/xiè/	/hé/	/hè/	/léi/	/lèi/
English words	Hair	Hat	Shoe	Crab	River	Drink	Thunder	Tears
Tone 3 - Tone 4								
Score (%)	92.5		80.0		75.0		73.8	
Chinese character	水	睡	脸	链	耳	二	鼠	树
Pinyin	/shuǐ/	/shuì/	/liǎn/	/liàn/	/ěr/	/èr/	/shǔ/	/shù/
English words	Water	Sleep	Face	Chain	Ear	Two	Mouse	Tree

Note: The identification score (%) shown for each minimal pairs was the average between the two words of a minimal pair.

líu lián/ (durian) were adopted after consultation with a Mandarin-speaking linguist. For Category 3 (Vowel Perception Test – VPT), a total of 16 words were tested and 12 words were selected. For Category 4 (Consonant Perception Test – CPT), a total of 21 words were tested and 12 words were selected.

Table 3 shows the words that were selected to form Category 5 (Tone Perception Test – TPT). There are six subcategories based on the tonal contrasts in Mandarin (i.e., Tone 1-Tone 2, Tone 1-Tone 3, Tone 1-Tone 4, Tone 2-Tone 3, Tone 2-Tone 4 and Tone 3-Tone 4 contrasts). Four minimal pairs are required to form each subcategory. Therefore, 31 minimal pair of words were tested and subsequently 24 pairs were selected.

In summary, a total of 80 words were selected to form the Malaysian Mandarin paediatric speech perception test. The selected words comprised of 57 monosyllabic words, 20 disyllabic words and three trisyllabic words. Of these 80 words, 66 words occurred once in the test whereas 14 words occurred more than once (12 words X 2 times = 24 test items, 2 words X 3 times = 6 test items) in the test. Therefore, the Malaysian Mandarin paediatric speech perception test has 96 test items as a whole.

PHASE II: DEVELOPMENT OF DIGITIZED WORD TOKENS AND SUITABILITY OF PICTURES

ACOUSTIC ANALYSIS AND VISUAL INSPECTIONS

A total of 480 word tokens were digitally recorded from the two talkers (80 words X three repetitions X two talkers). These word tokens were extracted from the carrier phrase and the duration of the extracted word tokens were documented. The average duration of female monosyllabic tokens was 0.44 seconds (SD = 0.11 seconds, range = 0.46 seconds). For female disyllabic tokens, the average duration was 0.76 seconds (SD = 0.09 seconds, range = 0.39 seconds) while for female trisyllabic tokens, the average duration was 1.04 seconds (SD = 0.07 seconds, range = 0.20 seconds). For male tokens, the average duration for monosyllables was 0.45 seconds (SD = 0.08 seconds, range = 0.36 seconds). For disyllabic tokens, the average duration was 0.73 seconds (SD = 0.08 seconds, range = 0.28 seconds), and for trisyllabic tokens, the duration was averaged at 0.80 seconds (SD = 0.05 seconds, range = 0.17 seconds). The reliability of the duration for the extracted word tokens was examined via Intraclass Correlation Coefficient (ICC). The ICC value of 0.95 was obtained, indicating excellent reliability (> 0.90; Koo & Li 2016) between the two testers.

The extracted 480 word tokens underwent acoustic analysis and visual inspections. Figure 1 displays the examples of spectrograms with pitch contours. Panels in the first row show examples of pitch contours for tokens of the same words with Tone 1, followed by Tone 2, Tone 3, and Tone 4 in subsequent rows. The left panels of Figure 1 demonstrate tokens that fulfilled the pitch contours criteria of Mandarin tones (Liu et al. 2007) whereas the right panels show pitch contours that were not acceptable. A total of 237 (49.4%) tokens passed the acoustic analysis and visual inspections. The remaining 243 tokens (50.6%) contained background noise, high-pitched hum or click sounds. Modifications (e.g., noise and click removal) were applied on these remaining tokens and 94 of the tokens were discarded post modification because they still did not fulfil the acoustic analysis criteria. Overall, 386 word tokens underwent the subsequent sound quality evaluation.

SOUND QUALITY EVALUATION

A total of 305 (79.0%) of the 386 tokens were rated as having good sound quality by two professionals. Disagreement occurred on 75 tokens (19.4%) which were then rated by a third professional. Out of these 75 tokens, 71 tokens were rated as having good sound quality and four tokens were rated as having poor sound quality. The remaining six (1.6%) of the 386 tokens which had poor sound quality were discarded. Since there was no token to represent the words 土 /tǔ/ (soil), 鼓 /gǔ/ (drum) and 眼 /yǎn/ (eye), re-recording (three tokens for each word) and acoustic analysis were conducted on these new tokens. In total, 384 tokens were included for subsequent validation.

SUITABILITY OF PICTURES AND VALIDATION OF DIGITALLY RECORDED WORD TOKENS

Picture-naming test was done with 77 pictures and 75 of the pictures obtained 100% correct naming score. Another two pictures that represented 土 /tǔ/ (soil) and 站 /zhàn/ (stand) achieved a score of 45% and 5%, respectively. Since the average score was 98% and was set as a benchmark to select pictures, the two aforementioned pictures were modified according to participants' feedback as shown in Figure 2. The modified pictures were then re-evaluated by the participants and a correct score of 100% was attained for both pictures.

Validation of the 384 tokens was conducted using speech identification according to test categories. The

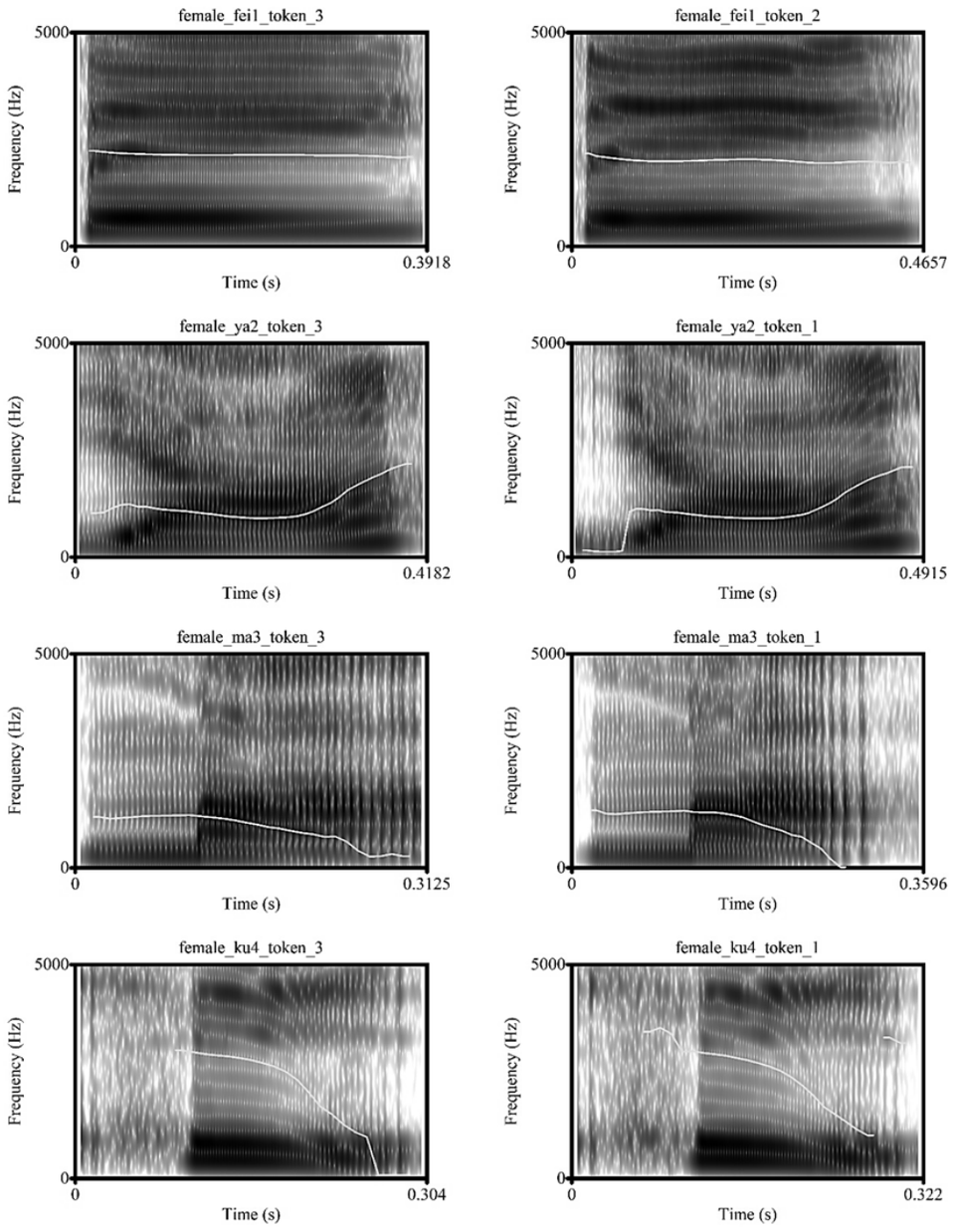

FIGURE 1: Examples of word spectrograms with pitch contours. The horizontal lines (either flat or slanted) represents the pitch contour of Mandarin tones. The figures on the left panel fulfilled (accepted) the Mandarin tones criteria. The figures on the right panel did not fulfil (not accepted) the Mandarin tones criteria.

number of tokens tested and selected (with the respective percentages) was portrayed according to talker gender and test categories in Table IV. All of the selected tokens achieved 100% identification score except 鱼 /yú/ (fish) in Category 3 which had 95% identification score. In addition, 鼻 /bí/ (nose), 图 /tú/ (picture) and 裤 /kù/ (trousers) in Category 5 were also scored at 95% identification score. A total of 192 tokens were selected as the finalized digitized test items for the five different test categories.

DISCUSSION

PHASE I: SELECTION OF TEST ITEMS

One of the selection criteria for the finalized test items was high familiarity words (Fu et al. 2011; Li et al. 2017; Umat et al. 2010; Zheng et al. 2009; Zhu et al. 2012). By seeking high familiarity words among the target population, face validity of the newly developed test can be established. This is important in order to ensure that the developed speech

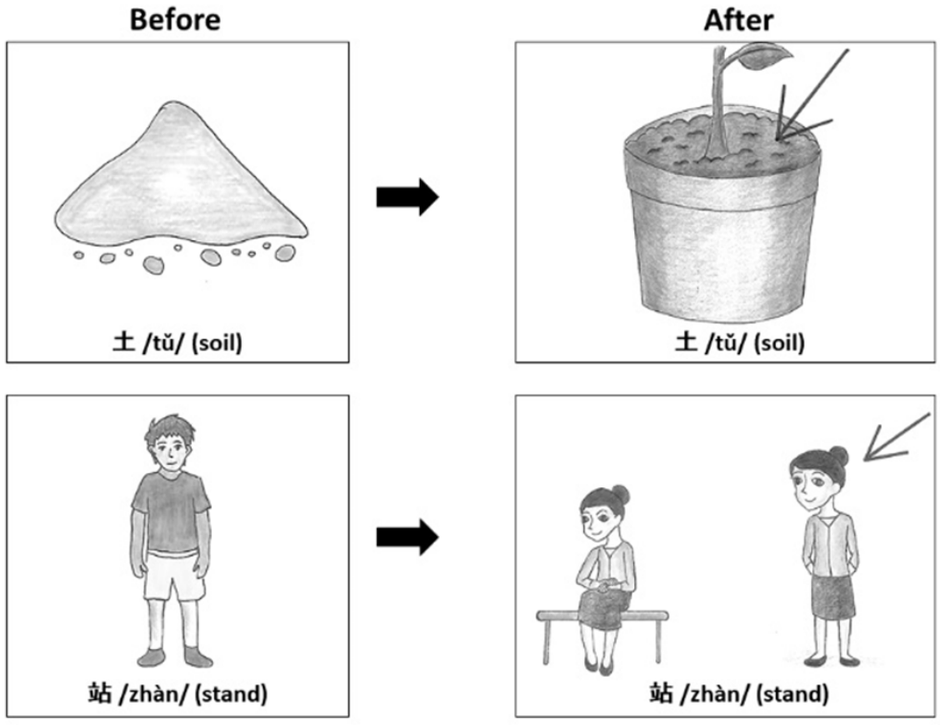

FIGURE 2: Picture modification on two test item pictures

TABLE 4. Number of word tokens tested and selected and the percentages of tokens selected (%; in parentheses) across test categories according to talker gender

Test Categories	No. of Tokens Tested		No. of Tokens Selected	
	Female Tokens	Male Tokens	Female Tokens	Male Tokens
Category 1 (SpPPT)	31	32	12 (38.7)	12 (37.5)
Category 2 (SPT)	30	28	12 (40.0)	12 (42.9)
Category 3 (VPT)	28	29	12 (42.9)	12 (41.4)
Category 4 (CPT)	31	29	12 (38.7)	12 (41.4)
Category 5 (TPT)	76	70	48 (63.2)	48 (68.6)
Total	196	188	96 (49.0)	96 (51.1)

perception test is able to reflect a child's true performance (Young & Kirk 2016) and to reduce the error rate in tone perception test which is closely related to the familiarity of the words (Zhu et al. 2014). In the current study, words were chosen from the vocabulary inventory among children aged two years old and tested among children of the same age in Phase I. This helps to ensure that the words selected were appropriate and fully mastered by the target population (Umat et al. 2010). Words with the highest identification scores in Phase I were determined as high familiarity words among the tested children. Most of these high familiarity words were nouns which represent common objects in daily life and verbs representing familiar activities. These words were also able to be represented in the form of coloured illustrations. Additionally, the current study also omitted words which are less frequently known in their monosyllabic pattern. The omitted words were 象 /xiàng/ (elephant), 椅 /yǐ/ (chair) and 兔 /tù/ (rabbit) which are more commonly known in their disyllabic forms which are 大象 /xiàng/ (elephant), 椅子 /yǐ zi/ (chair) and 兔子/tù zi/ (rabbit) respectively. These criteria are consistent with that of Umat et al. (2010) study.

Word familiarity among the children was assessed with a closed-set picture-pointing task (Peng et al. 2004; Umat et al. 2010; Zhu et al. 2014). The picture-pointing task is preferred over picture-naming task among young children as the children's speech and language skills are still developing and they may have accompanying errors

of speech (Markides 1987). Test items represented by coloured illustrations were also found to be suitable in successfully eliciting responses from young children aged as young as 30 to 35 months old in a speech perception test (Hodgson 1994; Zheng et al. 2009; Zhu et al. 2014). Therefore, the combination of high familiarity words with suitable illustrations and a picture-pointing task makes speech perception testing feasible among paediatric population with hearing loss or difficulty to articulate a response accurately (Zhu et al. 2014).

In order to create a more comprehensive speech perception test for Mandarin-speaking children in Malaysia, this study also incorporated the test item selection criteria similar to that of the MESP test (Zheng et al. 2009). Firstly, disyllabic words carrying Tone 3 were excluded from Category 2 (spondee perception test). According to the Mandarin tone sandhi rules (Chien et al. 2015; Zhu et al. 2014), the first syllable of a Tone 3 Mandarin spondee will be pronounced as Tone 2 rather than Tone 3. Therefore, disyllabic words with Tone 3 were excluded in Category 2 to avoid this tone alteration and to maintain phonetic balance of the tone (Wang et al. 2007). Secondly, the words within each subcategory must have the same initial consonant and tone for Category 3 (vowel perception test). Within the contrasts, words within each subcategory for Category 4 (consonant perception test) must have the same coda and tone. Due to this constraint, only one set of words were tested and selected for the Tone 1, Tone 2, and Tone 3 subcategories. For the Tone 4 subcategory, two sets of words were tested and the set of words with the highest identification score were selected. Thirdly, Zheng et al. (2009) showed that Category 5 (tone perception category) was the most challenging for the younger children as compared to older children. Therefore, minimal word pairs for constructing Category 5 (tone perception test) in the current study were selected based on the identification scores among the two- and three-year- old children rather than the average group score.

PHASE II: DEVELOPMENT OF DIGITIZED WORD TOKENS AND SUITABILITY OF PICTURES

Phase II was conducted to develop and validate the recorded test stimuli and test item pictures. The extraction of tokens involved subjective evaluation in determining the word boundaries in spectrograms (Dickinson et al. 2013; Tsiartas et al. 2009). Since the token extraction was mostly performed by one author, the reliability of the extracted tokens is vital to be examined to ensure that they only contained target words. Therefore, a second author extracted 20% of the tokens as a reliability examination. From the results, it was demonstrated that the duration of tokens extracted from the female and male talkers was comparable. This signified that the selection of word boundaries was consistent and the speech rate of the two talkers was similar for the same words. This was further supported by the ICC result where an excellent reliability was achieved for the word boundaries selection between the two testers.

Acoustic analysis is pivotal to be performed via spectrogram analysis when developing in-house digitized materials. For this analysis, auditory and visual inspections were conducted simultaneously so that subjective listening judgment was accompanied by a thorough visual analysis to examine word tokens. This helps to ensure that the word tokens were free of idiosyncrasies such as deviated pitch contour and undesirable extraneous noise (Cheesman & Jamieson 1996). These elements could negatively affect the quality of word tokens. In addition, for the pitch contour inspected via spectrograms, it was observed that Tone 3 and Tone 4 are carrying a similar falling contour shape. This finding is consistent with previous studies (Lim et al. 2015; Lim 2018) demonstrating that the pitch contour for Tone 3 is mid-low to low (2-1) for the Malaysian Mandarin (Maldarin), rather than mid-low, low to mid-high (2-1-4) for the Chinese Putonghua. Regardless of this similarity in terms of pitch contour, the Tone 3 word tokens with such pitch contours were acceptable and not discarded. Picture suitability evaluation was to ensure good representability of the test items (Zhu et al. 2014). The pictures representing 土 /tǔ/ (soil) and 站 /zhàn/ (stand) were modified as the former resembled "mountain" or "rock" while the latter resembled "boy" or "human" according to the participants' feedback. The picture of 站 /zhàn/ (stand) was modified to include two persons in order to create a contrast to the action of standing from sitting. The attainment of 100% correct score post modification suggested that the modified pictures possessed good representability of the respective test item.

For the validation of the recorded word tokens, the achievement of low identification scores for two words in particular, may be a result of the effect of spoken colloquial Malaysian Mandarin. The word 裤 /kù/ (trousers) was identified incorrectly by some participants as 哭 /kū/ (cry). This is because the latter Tone 1 word is colloquially pronounced as Tone 4, which is similar to 裤 /kù/ (trousers), in the context of Malaysian spoken Mandarin which has the Southern Mandarin accent (Hays 2015). As such, the average identification scores of these two words were

relatively lower than all the other words for both talker gender. Nevertheless, these words were still selected for the test due to limited test words. Therefore, it is vital to ensure that the children are well-conditioned and are familiarized with the tone of these words before administering the test in the future.

LIMITATIONS OF STUDY

There were a few limitations in the current study. One limitation in Phase I was the use of live voice during field test. This may cause the intensity and clarity of authors' voice difficult to be controlled. This shortcoming could be overcome by using a sound level meter to monitor the presentation of live voice. Phase II of the current study had several limitations. Firstly, the word tokens were recorded from student volunteers. Due to the lack of experience as voice talent, the talkers required a longer recording session as they made occasional mispronunciation. For future studies, it is suggested that talker selection could be done with a panel of judges to evaluate the vocal quality, standard dialect and pronunciation of the voice talent. Preferably, professional voice talents are to be hired. Secondly, background noise was present in the recorded word tokens because the recording was conducted in a semi-professional recording studio. Therefore, as mentioned earlier, a detailed acoustic analysis and word token modification are essential to develop good quality word tokens prior to subsequent evaluation. Recommendation for future studies is to perform recording in a studio with double-walled sound-tested booth (Cheesman & Jamieson 1996) to minimize the background noise. Thirdly, the number of word tokens was limited because some test items had all the tokens eliminated after the acoustic analysis and sound quality evaluation. It is recommended to increase the number of word repetitions during recording, so as to increase the number of word tokens available for selection at the later stage.

CLINICAL IMPLICATIONS

The methodology of the current study may serve as a guide for researchers to develop other digitized speech perception test materials. In addition, the test materials could be potentially used to evaluate Mandarin-speaking children in multi-ethnic multilingual country such as Malaysia and Singapore. This is because the majority Malaysian and Singaporean Chinese are descendants of Chinese from the Southern Chinese provinces in Mainland China and use the Southern Mandarin accent in their spoken Mandarin (Hays 2015).

CONCLUSION

In conclusion, materials for the Malaysian Mandarin paediatric speech perception test (MyMaPS) were developed. The selected test items are high familiarity words which are suitable for children aged three to six years old in Malaysia. The test item pictures are good representations of all test items in MyMaPS. The digitized test stimuli underwent several stages of evaluations and were validated to ensure that they have good acoustic quality. This study showed that objective and subjective evaluations are important in order to select the best digitized test items and pictures for developing a new speech perception test. Future research includes collecting normative data for MyMaPS and evaluating the test application in local audiology clinics.

ACKNOWLEDGEMENT

The authors would like to thank all participants and their parents for allowing their children to participate in the study. We would also like to thank Ms. Tong Lee Choo and Mr. Foong Jia Hao for their assistance during the recording sessions. Special gratitude is also due to Mr. Steven Lee Onn Wah and Ms. Tey Shi Rou for their contribution during sound quality evaluation process. Part of the studies were funded by the *Geran Gerakan Penyelidik Muda* (GGPM-2017-053) research grant from Universiti Kebangsaan Malaysia (UKM).

REFERENCES

Ang, A. L. 2010. Pembinaan ujian persepsi pertuturan bahasa mandarin untuk kanak-kanak berbangsa cina yang bertutur dalam bahasa mandarin di Malaysia. Bachelor Degree Thesis, Faculty of Health Sciences, Universiti Kebangsaan Malaysia.

Bess, F. H., Dodd-Murphy, J. & Parker, R. A. 1998. Children with minimal sensorineural hearing loss: prevalence, educational performance, and functional status. *Ear and Hearing* 19(5): 339-354.

Cheesman, M. F. & Jamieson, D. G. 1996. Development, evaluation and scoring of a nonsense word test suitable for use with speakers of Canadian English. *Canadian Acoustics* 24(1): 3-11.

Chien, Y., Sereno, J. A. & Zhang, J. 2015. Priming the representation of Mandarin tone 3 sandhi words. *Language, Cognition and Neuroscience* 31(2): 1-11.

Chok, S. S. 2001. First word in Mandarin acquired by the Chinese children in Malaysia. Bachelor Degree Thesis, Faculty of Health Sciences, Universiti Kebangsaan Malaysia.

Chong, F. Y., Cheoy, L. P., Mazlan, R., & Maamor, N. (2020). Performance-intensity functions of

Mandarin fricative-affricate nonsense word test: preliminary findings. Speech, Language and Hearing, 23(3), 121 – 132. DOI:10.1080/2050571X.2019.1576364

Chong, F. Y., Lee, O. W., Abdol, N. & Mazlan, R. 2018. Development of Mandarin fricative-affricate nonsense word test: Part I. selection of best exemplars. *Malaysian Journal of Health Sciences* 16: 179-185.

Department of Statistics Malaysia. 2018. Current population estimates, Malaysia, 2017-2018. https://www.dosm.gov.my/v1/index.php?r=column/cthemeByCat&cat=155&bul_id=c1pqTnFjb29HSnNYNUpiTmNWZHArdz09&menu_id=L0pheU43NWJwRWVSZklWdzQ4TlhUUT09 [3 September 2018].

Dickinson, M., Brew, C. & Meurers, D. 2013. *Language and Computers*. West Sussex: John Wiley & Sons, Ltd.

Eisenberg, L. S., Johnson, K. C. & Martinez, A. S. 2005. Clinical assessment of speech perception for infants and toddlers. https://www.audiologyonline.com/articles/clinical-assessment-speech-perception-for-1016 [30 July 2018].

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.

Hays, J. 2015. Chinese in Malaysia. http://factsanddetails.com/asian/cat66/sub418/entry-4307.html [15 April 2020].

Hendrayani. 2015. Malaysia Language Education Policy. Dlm Tochon, F. V. (pnyt.). *Language Education Policy Studies*. Madison: University of Wisconsin-Madison. http://www.languageeducationpolicy.org [24 September 2017].

Hodgson, W. R. 1994. Evaluating infants and young children. Dlm Katz, J. (pnyt.). *Handbook of Clinical Audiology*. Edisi ke-4, hlm. 465-475. Baltimore: Williams and Wilkins.

Koo, T. K. & Li, M. Y. 2016. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine* 15(2): 155-163.

Leitner, G., Hashim, A. & Wolf, H. G. (pnyt.). 2016. *Communicating with Asia: the future of English as a global language,* hlm. 188-204. Cambridge: Cambridge University Press.

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.

Lim, H. W. 2010. Phonological acquisition in three languages: a cross-sectional study in English-Mandarin and Malay. Tesis Dr. Fal, University of Sheffield, England.

Lim, H. W. 2018. Multilingual English-Mandarin-Malay phonological error patterns: an initial cross-sectional study of 2 to 4 years old Malaysian Chinese children. *Clinical Linguistics and Phonetics* 32(10): 889-912.

Lim, H. W., Wells, B. & Howard, S. 2015. Rate of multilingual phonological acquisition: evidence from a cross-sectional study of English-Mandarin-Malay. *Clinical Linguistics & Phonetics* 29(11): 793-811.

Liu, H., Tsao, F. & Kuhl, P. K. 2007. Acoustic analysis of lexical tone in Mandarin infant-directed speech. *Developmental Psychology* 43(4): 912-917.

Marinova-Todd, S. H., Siu, C. K. & Jenstad, L. M. 2011. Speech audiometry with non-native English speakers: the use of digits and Cantonese words as stimuli. *Canadian Journal of Speech-Language Pathology and Audiology* 35(3): 220-227.

Markides, A. 1987. Speech tests of hearing for children. Dlm. Martin, M. (pnyt.). *Speech Audiometry*, hlm. 155-170. London: Whurr Publishers Ltd.

Mendel, L. L. 2008. Current considerations in pediatric speech audiometry. *International Journal of Audiology* 47(9): 546-553.

Mendel, L. L. 2009. Subjective and objective measures of hearing aid outcome. http://www.audiologyonline.com/articles/subjective-and-objective-measures-hearing-891 [20 September 2016].

Mendel, L. L. & Danhauer, J. L. 1997. *Audiologic evaluation and management and speech perception assessment*. San Diego: Singular Publishing Group, Inc.

Millett, P. 2012. Understanding how well your child hears with hearing aids. http://successforkidswithhearingloss.com/understanding-how-well-your-child-hears-with-hearing-aids/ [5 Oktober 2016].

Nelson, H. D., Bougatsos, C. & Nygren, P. 2008. Universal newborn hearing screening: systematic review to update the 2001 US preventive services task force recommendation. *Pediatrics* 122(1): e266-e276.

Peng, S. -C., Tomblin, J. B., Cheung, H., Lin, Y. -S. & Wang, L. -S. 2004. Perception and production of Mandarin tones in prelingually deaf children with cochlear implants. *Ear and Hearing* 25(3): 251-264.

Ramkissoon, I. 2001. Speech recognition threshold for multilingual populations. *Communication Disorders Quarterly* 22(3): 158-162.

Theunissen, S. C., Rieffe, C., Soede, W., Briaire, J. J. & Ketelaar, L. 2015. Symptoms of psychopathology in hearing-impaired children. *Ear and Hearing* 36(4): e190.

Tsiartas, A., Ghosh, P. K., Georgiou, P. & Narayanan, S. 2009. Robust word boundary detection in spontaneous speech using acoustic and lexical cues. International Conference on Acoustics, Speech and Signal Processing. Anjuran IEEE Signal Processing Society. Taipei, Taiwan, April 19-24.

Umat, C., Chong, S. L. & Mukari, S. Z. M. 2010. Mandarin speech perception tests for Malaysian Chinese children. *Malaysian Journal of Health Sciences* 8(1): 31-37.

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(12): 719-731.

Young, N. M. & Kirk, K. I. (pnyt.). 2016. *Pediatric cochlear implantation: learning and the brain.* New York: Springer.

Zheng, Y., Meng, Z. L., Wang, K., Tao, Y., Xu, K. & Soli, S. D. 2009. Development of the Mandarin early speech perception test: children with normal hearing and the effects of dialect exposure. *Ear and Hearing* 30(5): 600-612.

Zhu, M., Wang, X. & Fu, Q. -J. 2012. Development and validation of the Mandarin disyllable recognition test. *Acta Oto-Laryngologica* 132(8): 855-861.

Zhu, S., Wong, L. L. N. & Chen, F. 2014. Development and validation of a new Mandarin tone identification test. *International Journal of Pediatric Otorhinolaryngology* 78(12): 2174-2182.

Elizabeth Shu Jun Lim
Foong Yen Chong
Tian Kar Quar
Cila Umat
Lee Lee Soh
Yee Ven Leu

Audiology Programme
Centre for Rehabilitation and Special Needs Studies
Faculty of Health Sciences
Universiti Kebangsaan Malaysia

Hui Woan Lim
Speech Sciences Programme
Centre for Rehabilitation and Special Needs Studies
Faculty of Health Sciences
Universiti Kebangsaan Malaysia

Corresponding author: Foong Yen Chong
Email: foongyen.chong@ukm.edu.my