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 finalised 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 berhubung keratan rentas. Fasa I melibatkan pengumpulan perbentukkan Mandarin (n=113) dengan gambar item ujian dan seterusnya diniilai ke atas 40 kanak-kanak penutur bahasa Mandarin (berumur 2;0 hingga 5;11 tahan) yang berpendengaran normal. Fasa II merangkumi rakaman digital perbentukkan 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 perbentukkan 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 diniilai 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 袜子 (wà zi/ 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 (UKM.1.21.3/244/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>
<thead>
<tr>
<th>Category</th>
<th>Structure</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 (SpPPT)</td>
<td>Words which are grouped according to the number of syllable (monosyllables, disyllabic trochees, disyllabic spondees and trisyllables)</td>
</tr>
<tr>
<td>2 (SPT)</td>
<td>Disyllabic spondees which are divided into three subcategories which are grouped according to the syllables tone</td>
</tr>
<tr>
<td>3 (VPT)</td>
<td>Monosyllables which are divided into four subcategories with similar consonant and tone but different vowels</td>
</tr>
<tr>
<td>4 (CPT)</td>
<td>Monosyllables which are divided into four subcategories with similar vowel and tone but different initial consonants</td>
</tr>
<tr>
<td>5 (TPT)</td>
<td>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).</td>
</tr>
</tbody>
</table>

*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.
<table>
<thead>
<tr>
<th>Test Categories</th>
<th>High Familiarity Words with Correct Identification Score (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Category 1</strong></td>
<td>Monosyllables</td>
</tr>
<tr>
<td>Score (%)</td>
<td>97.5</td>
</tr>
<tr>
<td>Chinese character</td>
<td>Cat</td>
</tr>
<tr>
<td>Pinyin</td>
<td>/māo/</td>
</tr>
<tr>
<td>English words</td>
<td>Doctor</td>
</tr>
<tr>
<td>Score (%)</td>
<td>95.0</td>
</tr>
<tr>
<td>Chinese character</td>
<td>医生</td>
</tr>
<tr>
<td>Pinyin</td>
<td>/yī shēng/</td>
</tr>
<tr>
<td>English words</td>
<td>Doctor</td>
</tr>
<tr>
<td>Score (%)</td>
<td>97.5</td>
</tr>
<tr>
<td>Chinese character</td>
<td>医生</td>
</tr>
<tr>
<td>Pinyin</td>
<td>/yī shēng/</td>
</tr>
<tr>
<td>English words</td>
<td>Doctor</td>
</tr>
</tbody>
</table>
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 /liú lián/ (durian) were excluded from the test.

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

<table>
<thead>
<tr>
<th>Subcategories</th>
<th>Pair #1</th>
<th>Pair #2</th>
<th>Pair #3</th>
<th>Pair #4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tone 1 - Tone 2</td>
<td>Score (%)</td>
<td>88.8</td>
<td>86.3</td>
<td>82.5</td>
</tr>
<tr>
<td>Chinese character</td>
<td>窗/床</td>
<td>飞/肥</td>
<td>鸭/牙</td>
<td>汤/糖</td>
</tr>
<tr>
<td>Pinyin</td>
<td>/chuāng/</td>
<td>/chuáng/</td>
<td>/fēi/</td>
<td>/fēi/</td>
</tr>
<tr>
<td>English words</td>
<td>Window</td>
<td>Bed</td>
<td>Fly</td>
<td>Fat</td>
</tr>
<tr>
<td>Tone 1 - Tone 3</td>
<td>Score (%)</td>
<td>83.8</td>
<td>81.3</td>
<td>75.0</td>
</tr>
<tr>
<td>Chinese character</td>
<td>车/尺</td>
<td>冰/饼</td>
<td>烟/眼</td>
<td>妈/马</td>
</tr>
<tr>
<td>Pinyin</td>
<td>/chē/</td>
<td>/chě/</td>
<td>/bīng/</td>
<td>/bǐng/</td>
</tr>
<tr>
<td>English words</td>
<td>Car</td>
<td>Ruler</td>
<td>Ice</td>
<td>Biscuit</td>
</tr>
<tr>
<td>Tone 1 - Tone 4</td>
<td>Score (%)</td>
<td>92.5</td>
<td>87.5</td>
<td>78.8</td>
</tr>
<tr>
<td>Chinese character</td>
<td>眼睛/眼镜</td>
<td>猫/帽</td>
<td>哭/裤</td>
<td>书/树</td>
</tr>
<tr>
<td>Pinyin</td>
<td>/yǎn jīng/</td>
<td>/yǎn jìng/</td>
<td>/māo/</td>
<td>/mào/</td>
</tr>
<tr>
<td>English words</td>
<td>Eye</td>
<td>Spectacle</td>
<td>Cat</td>
<td>Hat</td>
</tr>
<tr>
<td>Tone 2 - Tone 3</td>
<td>Score (%)</td>
<td>88.8</td>
<td>88.8</td>
<td>67.5</td>
</tr>
<tr>
<td>Chinese character</td>
<td>鱼/雨</td>
<td>写/写</td>
<td>笔/图</td>
<td>土/土</td>
</tr>
<tr>
<td>Pinyin</td>
<td>/yú/</td>
<td>/yǔ/</td>
<td>/xiě/</td>
<td>/xiē/</td>
</tr>
<tr>
<td>English words</td>
<td>Fish</td>
<td>Rain</td>
<td>Shoes</td>
<td>Write</td>
</tr>
<tr>
<td>Tone 2 - Tone 4</td>
<td>Score (%)</td>
<td>75.0</td>
<td>72.5</td>
<td>70.0</td>
</tr>
<tr>
<td>Chinese character</td>
<td>毛/帽</td>
<td>鞋/蟹</td>
<td>河/喝</td>
<td>雷/泪</td>
</tr>
<tr>
<td>Pinyin</td>
<td>/máo/</td>
<td>/mào/</td>
<td>/xiè/</td>
<td>/xiē/</td>
</tr>
<tr>
<td>English words</td>
<td>Hair</td>
<td>Hat</td>
<td>Shoe</td>
<td>Crab</td>
</tr>
<tr>
<td>Tone 3 - Tone 4</td>
<td>Score (%)</td>
<td>92.5</td>
<td>80.0</td>
<td>75.0</td>
</tr>
<tr>
<td>Chinese character</td>
<td>水/睡</td>
<td>眼/脸</td>
<td>链/耳</td>
<td>二/鼠</td>
</tr>
<tr>
<td>Pinyin</td>
<td>/shuǐ/</td>
<td>/shuì/</td>
<td>/liǎn/</td>
<td>/liăn/</td>
</tr>
<tr>
<td>English words</td>
<td>Water</td>
<td>Sleep</td>
<td>Face</td>
<td>Chain</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Test Categories</th>
<th>No. of Tokens Tested</th>
<th>No. of Tokens Selected</th>
<th>Female Tokens</th>
<th>Male Tokens</th>
</tr>
</thead>
<tbody>
<tr>
<td>Category 1 (SpPPT)</td>
<td>31</td>
<td>32</td>
<td>12 (38.7)</td>
<td>12 (37.5)</td>
</tr>
<tr>
<td>Category 2 (SPT)</td>
<td>30</td>
<td>28</td>
<td>12 (40.0)</td>
<td>12 (42.9)</td>
</tr>
<tr>
<td>Category 3 (VPT)</td>
<td>28</td>
<td>29</td>
<td>12 (42.9)</td>
<td>12 (41.4)</td>
</tr>
<tr>
<td>Category 4 (CPT)</td>
<td>31</td>
<td>29</td>
<td>12 (38.7)</td>
<td>12 (41.4)</td>
</tr>
<tr>
<td>Category 5 (TPT)</td>
<td>76</td>
<td>70</td>
<td>48 (63.2)</td>
<td>48 (68.6)</td>
</tr>
<tr>
<td>Total</td>
<td>196</td>
<td>188</td>
<td>96 (49.0)</td>
<td>96 (51.1)</td>
</tr>
</tbody>
</table>
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 (Maladarin), 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 站 /zhàn/ (stand) and 站 /zhán/ (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 /zhán/ (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 (Hay 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
Omn 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.
Children with minimal sensorineural hearing loss:
prevalence, educational performance, and functional
Cheesman, M. F. & Jamieson, D. G. 1996. Development,
evaluation and scoring of a nonsense word test
suitable for use with speakers of Canadian English.
Chien, Y., Sereno, J. A. & Zhang, J. 2015. Priming the
representation of Mandarin tone 3 sandhi words.
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


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