Folia Phoniatrica et Logopaedica

Research Article

Folia Phoniatr Logop 2024;76:411–419
DOI: 10.1159/000534726

Received: May 10, 2023
Accepted: October 12, 2023
Published online: November 7, 2023

Applying Computer-Based Language Test to Young Children

Yunyi Hu[a, b] Kathy Yuet-Sheung Lee[a, b] Tammy Hui Mei Lau[a]
Wilson Shing Yu[a] Michael C.F. Tong[a, b] Iris H.-Y. Ng[a, b] Thomas Law[a, b]

[a]Department of Otorhinolaryngology, Head and Neck Surgery, Faculty of Medicine, The Chinese University of Hong Kong, Hong Kong, Hong Kong SAR; [b]The Institute of Human Communicative Research, The Chinese University of Hong Kong, Hong Kong, Hong Kong SAR

Keywords
Computer-based test · Paper-pencil test · Language assessment · Young children

Abstract
Introduction: This study aimed at exploring the feasibility of applying a computer-based language test to young children aged 2–4 years. ***Methods:*** Thirty-two Cantonese-speaking children, aged 2–4 years, were recruited from local kindergartens. All participants underwent an assessment using both the computer-based and paper-pencil versions of the Macau Cantonese Language Screening Scale for Preschool Children, following a crossover study design. A short break of 15–30 min was provided between the two assessments. The data were analysed at three levels: the overall test, subcategory, and individual item levels. At the overall test and subcategory levels, data were analysed using the paired samples *t*-test and ICC. At the item level, the percentage of agreement and Cohen's kappa value were selected to assess the agreement of the two test formats. ***Results:*** Excellent agreement was found for the overall test level, and good agreement was observed for four of the five subcategories. At the individual item level, 28 of the 35 items showed more than 80% agreement, and 16 items showed substantial to almost perfect agreement. ***Conclusion:*** These results suggest that the two test formats give similar total scores and subcategory scores for children aged 2–4. For children older than 2 years and 6 months, the agreement for matching items is as high as 83.68% (1,318/1,575). The computer-based test is thus highly recommended for this group of children. For children younger than 2 years and 6 months, a modified computer-based test is suggested to accommodate their needs.

Introduction

Recent studies suggested that approximately 7% of 4- to 5-year-old children in English-speaking countries have developmental language disorders [1]. In the UK, the rate of language disorder has been reported to range from 5% to 10% [2, 3]. In the USA, it has been estimated that 7.4% of 5-year-old children were diagnosed with developmental language disorder [1]. Poor language skills have a profound impact on children. Children with a language disorder at a young age are facing difficulties in their later language and academic development [4–8]. Language disorders are also associated with social, emotional, and

Yunyi Hu is the first author, and Kathy Yuet-Sheung Lee is the co-first author.

karger@karger.com
www.karger.com/fpl

Correspondence to:
Kathy Yuet-Sheung Lee, kathy-lee@cuhk.edu.hk

behavioural problems [9]. The impact of a language disorder may persist into adulthood [8].

Considering the high prevalence and life-long impact of language impairment, early identification of individuals through accurate assessments is highly recommended. Traditionally, the diagnosis was done using paper-pencil tests (hereafter PPTs), such as Clinical Evaluation of Language Fundamentals Fourth Edition [10] and Preschool Language Scale Fifth Edition [11], for the English-speaking population, and Hong Kong Cantonese Oral Language Assessment [12] for the local Cantonese-speaking population in Hong Kong. In recent years, computer-based tests (hereafter CBTs) have been developed given their advantages over paper-pencil counterparts. CBTs are more time-efficient for both examiner and child [13, 14], have a higher level of standardization [15], and can immediately generate scores and interpret the child's response [16, 17]. It has also improved children's access to speech-language pathology services via telehealth [9].

In addition to its advantages, CBT has also been verified to be comparable to PPT when applied to children aged 5–9 years. Haaf, Duncan, Skarakis-Doyle, Carew, and Kapitan [14] conducted a pioneer study that compared computer-assisted administration of the Peabody Picture Vocabulary Test-Revised (PPVT-R) [18] with the original paper-pencil version. No significant difference in test scores was reported between the groups, which supports the hypothesis that the CBT version is equivalent to the PPT version. In a more recent study, Waite, Theodoros, Russell, and Cahill [9] confirmed that the total raw scores and the scaled scores for each subtest of the Clinical Evaluation of Language Fundamentals Fourth Edition [10] were not significantly different between the computer-assisted and paper-pencil versions. Moreover, the validity and reliability of the two test formats were similar. Given the high reliability and validity and the other benefits of CBTs (e.g., time efficiency and access to remote clients), the CBT format was highly recommended for children older than 5 years old.

However, it remains unclear whether CBT can be applied to younger children. This study, therefore, aimed to compare CBT and PPT formats for assessing young children's language ability (i.e., children aged from 2 to 4 years). Data from this investigation will be a valuable addition to existing research findings of test format selection from a clinical perspective. The results of this study will assist in test format selection and future test development for assessing young children's language ability.

Methods

Participants

Thirty-two children (13 boys [41%], 19 girls [59%]) from Hong Kong, with a mean age of 2 years, 11 months (age range: 2 years, 2 months to 3 years, 11 months), were recruited. Eleven children fell within the age range of 2 years–2 years and 5 months, while nine children were between 2 years and 6 months and 2 years and 11 months old. Additionally, four participants were between 3 years old and 3 years and 5 months old, and eight participants fell within the age range of 3 years and 6 months to 3 years and 11 months. All participants were native speakers of Cantonese, which is the locally spoken Chinese language. Written informed consent was obtained from a parent or guardian of each child. The parents or guardians were asked to confirm via a questionnaire that the child did not have hearing impairment, cognitive impairment, or physical disability. Ethical approval for the study was obtained from the Joint Chinese University of Hong Kong-New Territories East Cluster Clinical Research Ethics Committee (Ref. No. 2016.685). The data reported in this paper were collected from January 30, 2018 to February 26, 2018. To ensure participant confidentiality, the authors of this study had no access to any information that could identify individual participants after data collection. All data were well protected and stored in accordance with relevant regulations and guidelines. Any identifying information, such as names or contact details, was removed from the data before analysis to further protect participant privacy. The participants were randomly assigned to two groups, using a crossover study design (Johnson, 2010). The first group ($N = 15$) was tested using the CBT first and then the PPT (C-P group), while the second group ($N = 17$) was tested using the PPT first and then the CPT (P-C group).

Test Materials

The Macau Cantonese Language Screening Scale for Preschool Children (MacCLASS-P) [19] was used to assess the participants' language ability. MacCLASS-P is a valid standardised assessment that aims to provide a comprehensive language assessment for children aged 2–4 years. Two versions of the test have been developed: a long form that has 118 items and a short form that has 38 items. The short-form version was used in this study, as it requires less time to complete (approximately 15 min) and, as such, reduces the possible pressure on the children. The 38 items of the MacCLASS-P could be divided into five subcategories, namely, receptive vocabulary, receptive grammar, expressive vocabulary, expressive grammar, and pragmatics. The participants were required to answer questions either by pointing at the correct visual stimulus or by directly speaking their answers aloud. The visual stimuli were presented in a picture book, and the auditory stimuli were presented with live voice by a test administrator.

The assessment kit consisted of a comprehensive user manual that specified the user qualifications, guidelines for administration and scoring, standardisation procedures, and psychometric properties. The internal consistency, as evaluated by Cronbach's alpha, was 0.91. The intra-rater reliability was 0.985, with an intra-class correlation coefficient (ICC) of 0.908–0.995. Construct validity was established using various approaches, including structural validity. The coefficients between different subcategories ranged from 0.621 to 0.829. The corrected coefficients between subcategories and the whole test ranged from 0.485 to 0.765. All correlation coefficients were significant ($p < 0.05$).

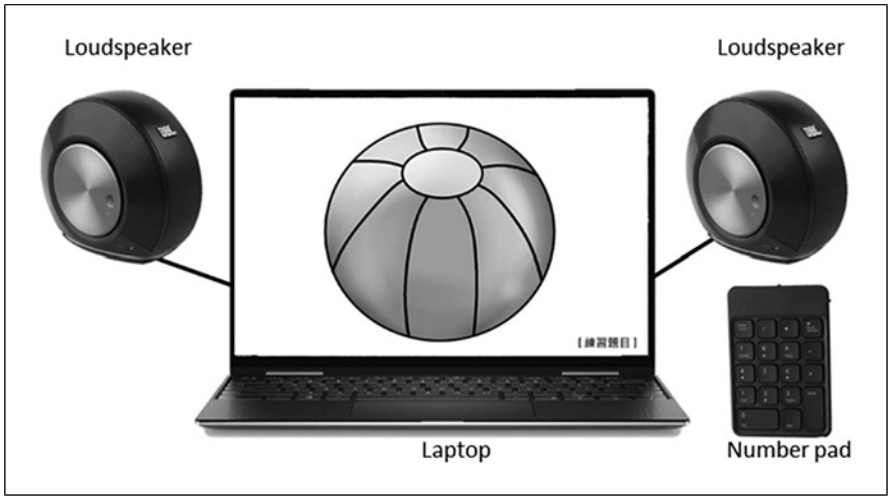

Fig. 1. Set-up of the computer version [20].

A computer version of MacCLASS-P that contained identical test items was developed by Lee, Lau, and Yu [20]. To allow the participants to familiarise themselves with the platform, two practice items were added to the original PPT items in the CBT. The computer version also takes approximately 15 min to complete, similar to the paper-pencil version. The only other difference between the computer version and the paper-pencil version was the presentation mode. In the computer version, the visual stimuli were presented on a laptop screen, while most of the auditory stimuli were presented via a recorded mode through loudspeakers. Figure 1 shows the set-up of the computer version, while the laptop screen shows an example of the practice items in this version.

However, three items in the pragmatic subcategory (e.g., calling the participant's name to see if he/she can respond) were solely conducted by the test administrator as they required real-time interactions. As the current study aims to investigate the comparability of CBT to PPT, these three items, which were not presented by the computer, will be excluded from further analysis. Their counterparts in the PPT version will also be excluded from further analysis. Thus, only 35 out of 38 items will be analysed in both versions. A summary of the test items that are selected for future analysis is listed in Table 1.

Test Procedures

Assessment Task

Each participant was assessed individually in a quiet room in kindergarten. One group was administered the CBT first, and the other group was administered the PPT first. Upon completion of the first test, a short break (15–30 min) was provided. All test items were administered following the procedures listed in the user manual. For both test conditions, the items were presented in the same order. Breaks during the tests were allowed when necessary. To minimise disruption to the classroom routine, each participant was taken out of the classroom once and assessed for a total of 30 min to 1 h.

CBT Administration Procedures

The CBT was administered using a laptop (model: Dell XPS 13 9360) with a screen size of 13 inches. The test program was installed on the laptop in advance. Two external loudspeakers (model: JBL Pebbles speaker) were connected to the laptop to play the audio recordings. A wireless number pad was also connected to the laptop for the administrator to control the program and assign scores ("0" for incorrect; "1" for correct). All the items could be repeated once if needed. After the participant responded, a score was immediately assigned by the test administrator, although changes could be made before a new item was presented. When the test was completed, an Excel report that listed the client's responses and scores for all of the items was generated for each participant.

PPT Administration Procedures

All auditory stimuli in the PPT version were presented using live voice by the test administrator. The corresponding visual stimuli were printed on a picture book and shown to the participants. After the participant responded, the test administrator recorded the score. The test administrator could also record the exact answer in addition to the score. When the test was completed, the test administrator calculated the scores manually.

Data Analysis

Data were analysed at three levels: the overall test, subcategory, and individual item levels. At the overall test and subcategory levels, data were analysed using the paired samples *t*-test and ICC. At the item level, the percentage of agreement and Cohen's kappa value were selected to assess the agreement of the two test formats. The SPSS statistical package (version 25; IBM, Armonk, NY, USA) was used for statistical analyses.

ICC is a reliability index that reflects both the degree of correlation and the agreement between measures [21]. In the current study, the two-way mixed-effects absolute model was selected [21] to examine the agreement between the CBT and PPT versions, at both total score level and subcategory level. In this case, the continuous quantitative variable was the raw score. The criteria for the ICC were set as ICC <0.5, poor agreement; ICC = 0.5–0.75, moderate agreement; ICC = 0.75–0.9, good agreement; and ICC >0.9, excellent agreement, based on the guidelines by Portney and Watkins [22].

The categorical variable, correct or incorrect, was used to calculate Cohen's kappa values. According to McHugh [23], the

Table 1. Examples of the selected items in each subcategory

Subcategory	Task type	Number of items	Example
Receptive language (17)			
Receptive vocabulary	Point at the correct visual stimulus	10	*Bin1 go3 hai6 biu1?** (Which picture is a "watch"?)
Receptive grammar	Point at the correct visual stimulus	7	*Bin1 fuk1 hai6 maa4maa1 mou5 syut3gou1?* (Which picture is describing "Mom does not have an ice-cream"?)
Expressive language (15)			
Expressive vocabulary	Say answers aloud	9	*Nei1 fuk1 hai6 me1 aa3?* (What is this picture?)
Expressive grammar	Say answers aloud	6	*Nei5 gin3 dou2 me1 aa3?* (What do you see here?)
Pragmatics (3)			
Pragmatics	Say answers aloud	3	*Nei5 gok3dak1 keoi5 dim2 aa3?* (How do you think he feels?)

*The Romanization system for Cantonese named Jyutping is used to transcribe the items.

level of agreement can be categorised into no (<0), slight (0–0.20), fair (0.21–0.40), moderate (0.41–0.60), substantial (0.61–0.80), or almost perfect (0.81–1.0) agreement.

Results

Overall Test and Subcategory Scores

The means, standard deviations, and 95% confidence intervals (CIs) of the difference between the PPT and CBT scores for each subcategory, as well as the overall scores, are presented in Table 2. A paired sample *t*-test was conducted to examine the level of agreement between the two test formats. No significant difference was found between the two test formats at the total score and subcategory level.

The subcategories of receptive language and expressive language were further divided into receptive vocabulary, receptive grammar, expressive vocabulary, and expressive grammar, respectively. The ICC values and their level of agreement are provided in Table 3.

Based on the ICC values, the agreement for overall scores between test formats was excellent (0.929). Good agreement was found for receptive grammar, expressive vocabulary, and expressive grammar.

Individual Item Scores

The percentage of agreement, Cohen's kappa values, and their level of agreement for individual items are shown in Table 4. At the individual item level, 28 of the 35 items shared more than 80% agreement, and 13 items showed substantial to almost perfect agreement.

Table 5 shows the distribution of matching responses and non-matching responses between the paper-pen format and the computer-based format for each age group. The percentages of matching responses (i.e., correct in both formats and incorrect in both formats) for the age groups ranged from 75.3% to 91.4%.

Discussion

This study aimed to explore the feasibility of applying the CBT to young children. Results from previous studies have compared the CBT and PPT on children aged 5–9. The results suggested that the PPT and CBT test formats do not differ significantly in terms of total scores [9, 13, 14, 24] or subcategory scores. This is consistent with the findings of the current study, which demonstrated good to excellent agreement between the two test formats and moderate to excellent agreement between the subcategories. An examination of the total scores suggested that the two test formats generated similar results in assessing the language ability of young children aged 2–4 years.

Though the results of the current study suggest that the CBT and PPT have a very good agreement for the total scores and subcategories when applied to children aged 2–4 years, the level of agreement at the individual item level should also be examined. Like the previous study by Waite et al. [9], which indicates that complete agreement at the individual item level was not achieved, the current study also confirmed the same conclusion. When being examined at the item level, the level of agreement between the two test formats was lower. In Waite et al. [9], the

Table 2. Comparison of the PPT and CBT scores

Subcategory (number of items)	Scores in the paper-pencil version		Scores in the computer-assisted version		95% CI of the difference		*t*	*p* value
	mean	SD	mean	SD	lower bound	higher bound		
Receptive language (17)	13.69	3.578	13.03	4.284	−1.489	0.176	−1.608	0.118
Expressive language (15)	9.47	4.080	10.16	3.878	−0.081	1.456	1.824	0.078
Pragmatics (3)	1.59	1.241	1.72	1.326	−0.130	0.380	1.000	0.325
Overall (35)	26.97	8.874	27.28	9.124	−0.837	1.462	0.272	0.778

Table 3. Agreement between the CBT and PPT based on the intra-class correlation for the overall test and subcategory levels

Subcategory	ICC value (single measures)	Level of agreement
Receptive language		
Receptive vocabulary	0.645	Moderate
Receptive grammar	0.817	Good
Expressive language		
Expressive vocabulary	0.787	Good
Expressive grammar	0.800	Good
Pragmatics	0.848	Good
Overall	0.929	Excellent

technical problems were considered to account for these discrepancies. For example, problems with the positioning of the participant in relation to the stimulus book, and with lighting and exposure to web cameras have caused difficulties for administrators to interpret the responses. The low speech volume and problems with the recording system were also considered to have affected the test results. However, Waite et al. [9] did not reckon children's behaviours have affected the results, unlike the previous study by Fairweather et al. [25] which found that younger children may need furniture modifications or may have difficulties in maintaining focus during the assessment.

In the current study, we agree with Fairweather and colleagues [25] that younger children's behavioural issues may have affected the results in CBT mode. Older children (2 years, 6 months to 3 years, 11 months) were able to achieve equally well in the CBT and PPT formats, with 83.68% (1318/1575) agreement for matching items. Though younger children could simultaneously look at the screen and listen to the speaker's voice during the CBT version, they displayed various behaviours while taking the test. Many children aged from 2 years to 2 years, 5 months looked at the loudspeaker and the number pad instead of the computer screen when the test started. Some children of this age showed curiosity towards the buttons on the computer keyboard and some of them touched the laptop keyboard and the keyboard buttons, expecting reinforcement feedback. Some of these children were also alarmed when they heard an unfamiliar voice from a loudspeaker with no corresponding face or facial expressions. In general, younger children were more familiar with face-to-face interactions and were more attentive during the PPT version. This behaviour contributed to a relatively low percentage agreement for matching items (i.e., CBT1-PP1 and CBT0-PP0), which fell below 80% in the youngest group. We speculate that differences in attention span and familiarity with the computer setup may have contributed to the discrepancies between the older groups and the youngest group. The behaviours displayed by the youngest group suggest a need to modify computer-based formats for this age group.

Using CBTs as a Clinical Option for Testing Young Children

CBTs have many merits that make them appealing to researchers and clinicians. Firstly, they provide a standardised presentation of test items [26], which is difficult to

Table 4. Agreement of individual items between the CBT and PPT

Subcategory	Item number	Percentage agreement (%)	Cohen's kappa	Level of agreement
Receptive vocabulary	RV-1	93.8	No data[#]	(Not included)
	RV-2	93.7	0.467*	Moderate
	RV-3	90.7	0.613*	Substantial
	RV-4	87.5	−0.049	NA[a]
	RV-5	87.5	0.273	NA[a]
	RV-6	84.4	0.664*	NA[a]
	RV-7	84.4	0.200	NA[a]
	RV-8	81.2	0.172	NA[a]
Receptive grammar	RG-1	87.5	0.636*	Substantial
	RG-2	84.4	0.600*	Moderate
	RG-3	84.4	0.646*	Substantial
	RG-4	81.2	0.600*	Moderate
	RG-5	75.0	0.382*	Fair
	RG-6	68.8	0.307	NA[a]
	RG-7	68.8	0.231	NA[a]
	RG-8	59.4	−0.030	NA[a]
	RG-9	59.4	0.148	NA[a]
Expressive vocabulary	EV-1	96.9	0.904*	Almost perfect
	EV-2	96.9	0.938*	Almost perfect
	EV-3	90.7	0.760*	Substantial
	EV-4	90.6	0.671*	Substantial
	EV-5	87.5	0.742*	Substantial
	EV-6	87.5	0.636*	Substantial
	EV-7	84.4	0.355*	Fair
	EV-8	68.8	0.310*	Fair
Expressive grammar	EG-1	96.9	0.840*	Almost perfect
	EG-2	90.7	0.529*	Moderate
	EG-3	87.5	0.634*	Substantial
	EG-4	84.4	0.675*	Substantial
	EG-5	84.4	0.518*	Moderate
	EG-6	65.6	0.200	NA[a]
	EG-7	62.5	0.216	NA[a]
Pragmatics	P-3	87.6	0.434*	Moderate
	P-4	84.4	0.683*	Substantial
	P-5	78.1	0.573*	Moderate

*$p < 0.05$. [#]No statistics were computed because one entry was a constant. [a]Level of agreement was not applicable because of an insignificant p value.

achieve in paper-pencil versions. Secondly, different randomised lists can be generated and implemented with greater ease in CBTs [26]. A randomised presentation of test items increases the reliability of the measure. To illustrate, the randomisation of test items reduces the possibility that items belonging to the same grammatical structure appear in succession. Thus, it can prevent the participants from receiving hints from the earlier test items. It can also prevent the participants from becoming bored or fatigued by continually answering questions related to the same type of item. Thus, the accuracy of the measurement of children's language ability is increased with randomisation. In summary, due to the advantages of standardisation and randomisation, CBTs increase the possibility of achieving more accurate measures of language ability for participants who are familiar with people-computer interaction.

Clinical Implications

The findings of this study have substantial clinical implications regarding the application of CBTs for evaluating language abilities in young children. The results of this study suggest that CBTs and PPTs are very similar for children older than 2 years and 6 months. The

Table 5. The distribution of matching responses and non-matching response between CBT and PPT in different age groups

Age group	No. of participants		CBT1 (%)	CBT0 (%)
2 years – 2 years, 5 months	11	PP1	38.2	14.0
		PP0	10.7	37.1
2 years, 5 months – 2 years, 11 months	9	PP1	62.5	8.6
		PP0	8.6	20.3
3 years – 3 years, 5 months	4	PP1	86.4	2.9
		PP0	5.7	5
3 years – 3 years, 11 months	8	PP1	62.1	8.6
		PP0	8.5	20.8
All	32	PP1	59.2	9.2
		PP0	8.7	22.8

degree of agreement on total scores, subcategory scores, and item scores was very similar, especially for language content areas within the same dimension. Given the numerous benefits associated with computer-based formats, it is highly recommended to prioritize the use of CBTs for children older than 2 years and 6 months.

However, it is crucial to underscore the significance of adapting the CBT to the specific needs and developmental abilities of young children, particularly those younger than 2 years and 6 months. Failing to implement appropriate modifications may potentially result in misdiagnosis or inaccurate assessment outcomes. Therefore, it is essential to exercise caution and ensure that the test is tailored to account for the unique characteristics, developmental milestones, individual needs, and engagement levels of younger children. Such considerations are vital to enhance the accuracy and validity of the assessment for this age group.

Future Directions of Test Modes

In addition to the aforementioned clinical implications, it is pertinent to explore future directions for test modes, specifically when evaluating the overall language ability of very young children (i.e., children under 2 years and 6 months). To accommodate the needs of these young children, it is worth investigating a modified CBT approach. One plausible adaptation involves employing a portable tablet instead of a complete computer set to present auditory and visual test stimuli. This modification aims to optimize the standardization of the test procedure while minimizing potential distractions for younger children, as tablets require fewer equipment pieces. Furthermore, the portability of tablets allows for greater flexibility in test administration, catering to individual needs. For instance, the test administrator can utilize the tablet to reinforce desired behaviour and encourage children to remain engaged, such as by inviting them to slide the screen to proceed to the next item.

However, it is important to note that the feasibility of implementing this modified CBT for young children, particularly those under 2 years and 6 months, warrants further investigation in future research. Subsequent studies can focus on assessing the practicality, effectiveness, and appropriateness of this adapted approach, taking into account specific developmental considerations and engagement strategies required for very young children. These future directions underscore the potential utilization of technology and the adaptation of test modes to better suit the needs of young children, ultimately enhancing the assessment process and bolstering the accuracy of evaluating their language abilities.

Limitations

This study has some limitations. Firstly, the inclusion criteria for participants were based solely on parent report, without further objective verification. This reliance on subjective information increases the potential for inaccuracies or misinterpretations regarding the absence of hearing impairment, cognitive impairment, or physical abnormalities in the children. As a result, there is a possibility of including participants who may have had undetected or unreported conditions (e.g., hearing loss due to otitis media), which could have influenced the study outcomes. Secondly, there was no inter-rater or intra-rater reliability test. However, in the experiment, responses to comprehension items (e.g., "*bin1 fuk1 hai6* '*ping4gwo2*'?," "Which one is the 'apple'?") involved pointing to the correct visual stimulus among six choices, and responses to expression items (e.g., "*nei1 fuk1 hai6 mat1je5?*", "What is this?") involved mostly delivering

simple answers, such as a noun or a short phrase. Given the nature of these responses, the reliability of the scoring is likely to be high. Third, it would be better to equal the sample sizes for the age groups and increase the number of participants for better generalization of the study findings. However, our results provide valuable insights that will inform our future work to finalize the format of this language-assessment tool.

Statement of Ethics

Ethical approval for the study was obtained from the Joint Chinese University of Hong Kong-New Territories East Cluster Clinical Research Ethics Committee (Ref. No. 2016.685). Written informed consent was obtained from a parent or guardian of each child.

Conflict of Interest Statement

The authors report no declarations of interest.

Funding Sources

Financial support for this study was provided by the Health and Medical Research Fund, Government of Hong Kong Special Administrative Region (Grant No. 14152301) and the Knowledge Transfer Project Fund, The Chinese University of Hong Kong (Grant No. KPF21GWP14). We would like to extend our gratitude towards all the participants and their families for their participation in this project. We are also thankful to the schools for their help in recruiting our participants.

Author Contributions

Yunyi Hu: conceptualization, data curation, writing – original draft, and visualization. Kathy Yuet-Sheung Lee: conceptualization, methodology, writing – reviewing and editing, and supervision. Tammy Hui Mei Lau: data curation, investigation, and writing – reviewing and editing; Wilson Shing Yu: conceptualization, methodology, data curation, investigation, and writing – reviewing and editing. Michael Chi Fai Tong: supervision, funding acquisition. Iris H.-Y. Ng: writing – reviewing and editing, supervision. Thomas Law: conceptualization, writing – reviewing and editing, and supervision.

Data Availability Statement

Data cannot be shared publicly since the Chinese University of Hong Kong owns the data. Data are available from the Chinese University of Hong Kong-New Territories East Cluster Clinical Research Ethics Committee (Ref. No. 2016.685) for researchers who meet the criteria for access to confidential data. All data generated or analysed during this study are included in this paper. Further enquiries can be directed to the corresponding author.

References

1 Calder SD, Brennan-Jones CG, Robinson M, Whitehouse A, Hill E. The prevalence of and potential risk factors for developmental language disorder at 10 years in the raine study. J Paediatr Child Health. 2022;58(11):2044–50.
2 Boyle J, Gillham B, Smith N. Screening for early language delay in the 18–36 month age-range: the predictive validity of tests of production, and implications for practice. Child Lang Teach Ther. 1996;12(2):113–27.
3 Norbury CF, Gooch D, Wray C, Baird G, Charman T, Simonoff E, et al. The impact of nonverbal ability on prevalence and clinical presentation of language disorder: evidence from a population study. J Child Psychol Psychiatry. 2016;57(11):1247–57.
4 Bishop DV, Adams C. A prospective study of the relationship between specific language impairment, phonological disorders and reading retardation. J Child Psychol Psychiatry. 1990;31(7):1027–50.
5 Clegg J, Hollis C, Mawhood L, Rutter M. Developmental language disorders: a follow-up in later adult life. Cognitive, language and psychosocial outcomes. J Child Psychol Psychiatry. 2005;46(2):128–49.
6 Johnson CJ, Beitchman JH, Young A, Escobar M, Atkinson L, Wilson B, et al. Fourteen-year follow-up of children with and without speech/language impairments: speech/language stability and outcomes. J Speech Lang Hear Res. 1999;42(3):744–60.
7 Stothard SE, Snowling MJ, Bishop DV, Chipchase BB, Kaplan CA. Language-impaired preschoolers: a follow-up into adolescence. J Speech Lang Hear Res. 1998;41(2):407–18.
8 Young AR, Beitchman JH, Johnson C, Douglas L, Atkinson L, Escobar M, et al. Young adult academic outcomes in a longitudinal sample of early identified language impaired and control children. J Child Psychol Psychiatry. 2002;43(5):635–45.
9 Waite MC, Theodoros DG, Russell TG, Cahill LM. Internet-based telehealth assessment of language using the CELF–4. Lang Speech Hear Serv Sch. 2010;41(4):445–58.
10 Semel E, Wiig EH, Secord WA. Clinical Evaluation of Language Fundamentals. 4 ed. Toronto, Canada: The Psychological Corporation/A Harcourt Assessment Company; 2003.
11 Zimmerman IL, Steiner VG, Pond RE. Preschool Language Scale–fifth edition (PLS-5). San Antonio: Psychological Corporation; 2011.
12 T'sou B, Lee T, Tung P, Man Y, Chan A, Tou CKS. Hong Kong Cantonese Oral Language Assessment Scale. Hong Kong: City University of Hong Kong; 2006.
13 Carson K, Gillon G, Boustead T. Computer-administrated versus paper-based assessment of school-entry phonological awareness ability. Asia Pac J Speech Lang Hear. 2011;14(2):85–101.
14 Haaf R, Duncan B, Skarakis-Doyle E, Carew M, Kapitan P. Computer-based language assessment software: the effects of presentation and response format. Lang Speech Hear Serv Sch. 1999;30(1):68–74.
15 Collerton J, Collerton D, Arai Y, Barrass K, Eccles M, Jagger C, et al. A comparison of computerized and pencil-and-paper tasks in assessing cognitive function in community-dwelling older people in the Newcastle 85+ Pilot Study. J Am Geriatr Soc. 2007;55(10):1630–5.

16 Björnsson JK. Changing Icelandic national testing from traditional paper and pencil based tests to computer based assessment: some background, challenges and problems to overcome. Towards a research agenda on computer-based assessment. Vol. 10; 2008.
17 Ripley M. Transformational computer-based testing. In: Scheuermann F, Björnsson J, eds. The transition to computer-based assessment: new approaches to skills assessment and implications for large-scale testing. Luxembourg: Office for Official Publications of the European Communities; 2009. p. 92–8.
18 Dunn LM, Dunn LM. Peabody Picture Vocabulary Test-revised. Circle Pines: American Guidance Service; 2007.
19 Deaf Association Macau. Manual on the Macau Cantonese Language Screening Scale for Preschool children (MacCLASS-P). Macao Special Administrative Region; 2016.
20 Lee KYS, Lau THM, Yu WS. Performance of pre-schoolers on two testing modes on a language assessment: Hong Kong Cantonese Language Assessment Scale for Preschool children (HK-CLASS-P). Hong Kong speech and hearing symposium. Hong Kong; 2018.
21 Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. J Chiropr Med. 2016;15(2):155–63.
22 Portney LG, Watkins MP. Foundations of clinical research: applications to practice. Upper Saddle River, NJ: Pearson/Prentice Hall; 2009.
23 McHugh ML. Interrater reliability: the kappa statistic. Biochem Med. 2012;22(3):276–82.
24 Maguire KB, Knobel ML, Knobel BL, Sedlacek LG, Piersel WC. Computer-adapted PPVT-R: a comparison between standard and modified versions within an elementary school population. Psychol Sch. 1991;28(3):199–205.
25 Fairweather C, Parkin M, Rozsa M. Speech and language assessment in school-aged children via videoconferencing. In: 26th World Congress of the International Association of Logopedics and Phoniatrics. Brisbane, Australia; 2004.
26 Inouye DK, Bunderson VC. Four generations of computerized test administration. Machine-Mediated Learn. 1986;1(4):355–71.