The development of Hong Kong norms for students' grammatical knowledge of written Chinese and its application in assessing the competence of deaf and hard-of-hearing (DHH) students

by

YIU, Kun Man Chris

A Thesis Submitted to
The Education University of Hong Kong
in Partial Fulfillment of the Requirement for
the Degree of Doctor of Education

18 September 2023

Statement of Originality

I, YIU, Kun Man Chris, hereby declare that I am the sole author of the thesis and the material presented in this thesis is my original work except those indicated in the acknowledgement. I further declare that I have followed the University's policies and regulations on Academic Honesty, Copyright and Plagiarism in writing the thesis and no material in this thesis has been submitted for a degree in this or other universities.

Abstract

With limited access to spoken or signed languages, many deaf or hard-of-hearing (DHH) struggle with language acquisition and literacy development regardless of their communication modality (Spencer & Marschark, 2010). For more than three centuries, the reading achievement of DHH students has been consistently lagging behind that of their hearing peers, a phenomenon being characterized as the "fourth-grade ceiling", i.e., high school DHH graduates show a consistent result of an average reading level of fourth grade or below (Babbidge, 1965; Qi & Mitchell, 2012) though more evidence has begun to show that DHH students are able to surpass this ceiling in their reading levels (Mayer, Trezek, & Hancock, 2021).

Grammatical knowledge of DHH students is considered an essential building block in reading development (Kelly 1996). However, DHH students' knowledge of morphosyntax involving functional categories, which function as an essential component of grammar, are extremely vulnerable. Consequently, students' difficulties in reading comprehension affect their academic performance (Spencer & Marschark, 2010). Teachers and speech therapists need validated assessment tools that help understand DHH students' grammatical development and design effective interventions to cater for students' individual needs (Cannon, et al., 2011).

The sociolinguistic context in Hong Kong is unique. While a great majority of children speak in Cantonese, the written language they use and learn at school is written Chinese, which follows the grammar of Mandarin Chinese (Wang, 2019). No tool is available for measuring HK Cantonese-speaking DHH children's grammatical knowledge in written Chinese. There are only a few oral language assessments in Hong Kong include some items specifically assessing Cantonese morphosyntax, for example, the Hong Kong Cantonese Oral Language

Assessment Scale (HKCOLAS) (T'sou et al., 2006) and the Hong Kong Test of Preschool Oral Language (TOPOL; Wong et al., 2019).

The study is part of a larger project that aims to develop a tool to measure the grammatical knowledge of DHH children in written Chinese. In this study, the psychometric properties of the original 172-item profiling tool, namely the Chinese Grammatical Assessment (CGA), was thoroughly reviewed through Rasch analysis based on a dataset with 963 typically hearing students and 40 deaf and hard of hearing students. An expert panel with ten subject matter experts (SMEs) were set up to conduct content validation for CGA. The representativeness of the grammatical categories, and the appropriateness and relevance of the test items of CGA were thoroughly reviewed by the SMEs.

Regarding the findings and recommendations through content validation and the psychometric review, alternate forms with 46 items were established to develop two CGA short tests. With further confirmation of the validity and reliability of CGA, the norms for the two CGA short tests were set up in percentile ranks and applied in a group of deaf and hard-of-hearing students as a case study, aims to further review the reliability and validity of the assessment. Finally, CGA scores collected from the two short tests were found highly correlated with the academic performance of both typically developing (TD) and DHH students. CGA scores can also significantly predict students' academic performance in Chinese Language.

Keywords: written Chinese, grammatical knowledge, Deaf and Hard-of-Hearing (DHH), assessment, standardization

Acknowledgement

I would like to take this opportunity to express my heartfelt gratitude to all the persons who have supported me all the way through this special journey. It would be impossible for me to complete this thesis without their unfailing support.

I owe my sincere thanks to my principal supervisor Dr. Kevin Yuen and my associate supervisor Dr. Anna Kam for their encouragement and support to my thesis writing. Here, I would also like to express my deepest gratitude to the Members of the Examination Panel, including Professor Liu Hsiu-Tan, Dr. Randolph Chan and Dr. Lee Kwai Sang. Their comments and valuable ideas inspired me deeply on my thesis and also my objectives and aspirations in future research.

I am also deeply indebted to Professor Gladys Tang for her continuous support to my professional development in linguistics and deaf education research. I am grateful to have this invaluable opportunity to participate in the development of the Chinese Grammatical Assessment (CGA) for deaf and hard-of-hearing students. Special thanks to the research team that has been supporting the development of the CGA from its onset for almost ten years. My research project would never be successful without the guidance from Professor Gladys Tang and the participations of our dearest working companions including Dr. Li Qun, Ms. Li Jia, Mr. David Lam, Mr. Kevin Yu and Ms. Ma Shuya, Mr. Timothy Chan, Ms. Ivy Liu, and Ms. Anna Pun. It is always a great honour to be a member of the Centre for Sign Linguistics and Deaf Studies and the SLCO Community Resources Limited during my research journey.

I especially want to give my special thanks to Dr. Iris Ng who always share with me her concrete experience in developing a newly established assessment tool. She has always given me the most practical tips and advice when I was stuck in making decisions on my research project.

Thanks to all the students and parents who are willing to support this research and participate in all different data collection procedures. I am especially thankful to Rev. Slawomir Kalisz, Principal Wong Chi-Chung and all the teachers and school staff of Oblate Primary School. Their full participation in the Sign Bilingualism and Co-enrollment in Deaf Education Programme has given deaf and hard-of-hearing students a completely different learning experience in Hong Kong.

I must express my sincere thanks to my lovely wife Ms. Heidi Yuen for her unconditional love to me, and her persistent endurance to my unintended negligence to her during this long process of thesis writing. My two sons, Shun and Hymn, always give me encouragement to keep on going when I am about to give up. Their words of support always reminds me of their love and care that builds me up and keeps me going.

Last but not least, I would love to dedicate this thesis to my dearest mother Ms. Ng Suk-Ying who is resting in God's mighty hands peacefully.

Table of Contents

List of Abbreviations

ABI	Auditory Brainstem Implant
CI	Cochlear Implant
CGA	Chinese Grammatical Assessment
CGA-A	Chinese Grammatical Assessment-Form A
CGA-B	Chinese Grammatical Assessment-Form B
CVI	Content Validity Index
DHH	Deaf and Hard-of-hearing
HA	Hearing Aid
MNSQ	Mean Square Values
P1, P2,…P6	Primary One, Primary Two,…Primary Six
PTMEA-CORR	Point Measure Correlation
SLCO	Sign Bilingualism and Co-enrollment in Education
SME	Subject Matter Expert
TD	Typically Developing
ZSTD	Z-standardized Values

List of Figures

List of Tables

Chapter 1: Introduction

1.1 Statement of Purpose

Grammatical knowledge is of utmost importance for children to develop a language. The delay in the development of grammatical knowledge in children has proved to be a long-term impact not only on their oral communication but also on their literacy, including reading and writing. Deaf and hard-of-hearing children have long been defined as a group of disadvantaged language users because of the hearing deprivation they acquired. Delayed language development and low literacy skills tremendously extend the impact on their whole-person development, including but not limited to their social (Van Gent, 2016), cognitive (Hall et al., 2017), and academic development (Qi & Mitchell, 2012). However, auditory deprivation should not necessarily lead to literacy and cognitive deficiencies. Empirical evidence shows that deaf and hard-of-hearing (DHH) children possess normal cognitive and intellectual potentials, however, their language competence, including oracy and literacy, is still inferior to their hearing peers. Some studies observed promising results in DHH children's vocabulary development (Duchesne, 2016). Grammatical development in DHH children, in terms of their knowledge of morphosyntax, is still a struggling area that requires lots of guidance and support from educational and speech and language professionals though there is evidence showing that early cochlear implantation is conducive to DHH primary school students' comprehension of morphosyntactic reading comprehension (López-Higes, et al., 2015).

As a general phenomenon in Hong Kong, children speak in Cantonese but write or read in written Chinese, which follows the grammar of Mandarin. The demand of handling a different set of grammar when developing literacy is a challenge to all typically developing children in Hong Kong, especially for the first few years of formal education when Chinese literacy becomes the focal medium of instruction in most local classrooms. With no exception, DHH children facing severe challenges in their oral language development in Cantonese also experience problems in Chinese literacy following a different set of grammar. Developing a grammatical assessment in written Chinese would be a significant achievement to help investigate how much they have developed their grammatical knowledge in written Chinese. Of course, the tool is also very useful in providing evidence that helps us better understand Cantonese-speaking DHH students' development pathway in written Chinese.

The current study aims at developing and validating an assessment tool for measuring the grammatical knowledge of Cantonese-speaking children in written Chinese. The assessment tool, namely the Chinese Grammatical Assessment (CGA), focuses on children's receptive grammar. No production tasks are included. Two alternate forms would be developed with local norms based on data collected from typically developing (TD) children studying in local primary schools. The assessment will also be evaluated for its validity and reliability for reviewing DHH children's Chinese grammatical development.

This chapter describes the aim, objectives and scope of the research as well as the structure of the thesis. In the following sections, for the examples in written Chinese are required, the gloss in English will be provided such as *蘋果* 'apple' and in some cases, especially for some function words that no direct meaning can be provided, a phonetic representation following Cantonese Jyutping romanization system (The Linguistic Society of Hong Kong, n.d) will be provided for reference, for example, *了*(*liu5*).

1.2 Background and Justification of the Study

Literacy could be narrowly defined as "the ability to decode print that facilitates the acquisition of world knowledge through reading, which involves the comprehension of successive words and sentences from the bottom up" (Takashi, et al., 2017, p.88). Of course, this definition puts aside writing or language production and pays more attention to the receptive side of literacy. From a 'bottom-up' model (Flesch, 1955), children's learning to read requires the mastery of decoding rules of a written language. In contrast, in a 'top-down' approach, reading is a whole-to-part process; decoding every letter or word is unnecessary (Smith, 1994). However, solely relying on world knowledge and contextual clues to read is never enough. As readers proceed with reading, it is a highly complex cognitive activity. As proposed by the interactive model, reading is not a unidirectional process. Instead, it is a constructive interaction combining bottom-up and top-down processes (Barr, Sadow, & Blachowicz, 2002). Under such an assumption, readers are in an integrated use of their world knowledge and the acquired decoding such as phonology, morphology, semantic, and syntactic skills to comprehend the written text.

No matter which approach we are adopting, effective reading requires 'not only knowledge of vocabulary, but also rules of combining words into grammatical constituents to form simple or complex sentences, as well as knowledge of discourse rules for comprehension and production of a coherent text (Tang et al., 2022). From a simple view of reading, adequate manipulation of decoding and linguistics comprehension is the basis for reading (Hoover and Gough, 1990). No matter which component is lacking, reading will not be successful.

The reading achievement of DHH students has been consistently lagging behind their hearing peers. This worldwide phenomenon is being characterized as the "fourth-grade ceiling", i.e., high school DHH graduates at 18 years old or above show a consistent average reading level of fourth grade or below (Babbidge, 1965; Holt, 1993; Traxler, 2000; Qi & Mitchell, 2012).

Delayed language development is often reported to be a factor that adversely affects the reading development and educational achievement of DHH children (Spencer & Marschark, 2010). With the tremendous advancements in hearing technologies such as digital hearing aids and cochlear implants, there is growing evidence to show that some groups of DHH students may soon surpass this fourth-grade ceiling (Mayer, Trezek, & Hancock, 2021). Archbold & Mayer (2012) expalined that cochlear implantation helps to alleviate DHH students' barriers to learning and communication in the classroom and enhance their academic potential, but the impact on their ultimate attainment still varies. Besides, proficient sign language skills are also found to be an effective predictor of reading comprehension for DHH students in a sign-bilingual school setting (Dammeyer, 2014; Scott & Hoffmeister, 2017). Bilingual education in both sign langauge and spoken langauge has been proposed to be a safety net for DHH students with notably diverse hearing and speech perception abilities. It contributes positively to DHH students' academic development in both special school (Lange, Lane-Outlaw, Lange, & Sherwood, 2013) and mainstreamed co-enrollment settings when sign language is included as one of the medium of instructions (see Marschark, Knoors, & Antia, & 2019). However, the majority of children with hearing loss continue to experience restricted or ineffective access to spoken or signed language and subsequently tremendous struggles with literacy regardless of their communication modality (Moeller et al., 2007; Lederberg, Schick, & Spencer, 2013).

1.3 Phonological versus Grammatical Knowledge

Many studies have attempted to unravel which component(s) of language knowledge predict literacy development especially reading, but so far, no simple conclusion can be made. Some studies highlighted the significance of phonological awareness on DHH children's literacy development (Harris & Beech, 1998; Easterbrooks, et al., 2008; James, et al., 2008). In

Mayberry, del Giudice, and Lieberman's (2011) study, phonological awareness explained 11% of the overall variance, much less than the 35% variance predicted by children's overall language ability in either signed or spoken language. Even when the DHH students are using advanced hearing technology like cochlear implant which can effectively enhance their auditory access to speech information (Lee, van Hasselt, & Tong, 2010), only 26% of the overall variance of DHH students' literacy development was contributed from phonological processing, much lower than the 47% variance contributed from the overall language ability including vocabulary and syntactic abilities (Geer, 2003).

The relationship between phonological coding or awareness and reading ability in a phonetic language like English may be very different from that in a logographic language like Chinese (Ku & Anderson, 2003; Tong et al., 2009; Ching & Nunes, 2015) because of their different writing systems. Enhanced morphological awareness substantially improved children's literacy measures in Chinese (Wu et al., 2009), but that does not mean that phonological awareness has no significance in Chinese reading, though the impact may be relatively less than that in English (Taylor, 2002). How DHH students acquire written Chinese should be a quite different pathway from that in English literacy. Some research explores how DHH students comprehend some structures of written Chinese (Lam, 2016; Wang, Lian, & Lin, 2018; Wang & Andrews, 2020; among others), but further studies are required.

Vocabulary knowledge is considered a factor affecting DHH children's reading performance (Brisbois, 1995; Yamashita, 1999; Qian, 2002). It also interacts significantly with DHH children's morphosyntactic or grammatical knowledge when their performance in reading comprehension is concerned (Kelly, 1996; Gaustad & Kelly, 2004). In a review of a set of studies on the development of grammatical competence by DHH children who received operation of cochlear implants as early as before age 2, Duchesne (2016) found that many children with good vocabulary knowledge still faced great difficulties in developing their grammatical competence, and the struggle could be extended to their adolescence. DHH children's knowledge of morphosyntax is extremely vulnerable, especially in the structures that involve functional categories, which are the relatively 'unstressed' components of the grammar (see Quigley et al. 1976; Wilbur, Goodhart & Montandon, 1983; Berent 1988, 1996; de Villiers, de Villiers & Hoban, 1994; Lillo-Martin 1998; Friedmann & Szterman, 2006, 2011; Volpato, 2010; Guasti et al., 2014; Yiu, 2004, 2012; Lam, 2015, and among others).

Grammatical knowledge is considered an essential building block of DHH children's reading development (Kelly, 1996), however, there are only limited studies investigating DHH children's acquisition of Cantonese grammar such as relative clauses (Yiu, 2004; Lam, 2015), passive constructions (Yiu, 2012), etc. In these studies, DHH children are observed to be experiencing significant difficulties in their development of grammatical structures when compared to their typically developing counterparts. According to Lau et al. (2019), 18 (18%) and 40 (41%) out of 98 DHH students in local primary schools were found to have mild-to-moderate and severe language impairment respectively. Among the six different testing components of the Hong Kong Cantonese Oral Language Assessment Scale (HKCOLAS) (T'sou et al., 2006), including "Hong Kong Cantonese Grammar", "Textual Comprehension", "Word Definition", "Lexical-semantic Relations", "Narrative" and "Expressive Nominal Vocabulary", DHH students exhibited the poorest performance in the subtest of Cantonese Grammar. The more severe the hearing loss, the more significant gap between the standard scores and their comprehension and production of Cantonese grammar (Lau et al., 2019).

1.4 Motivation for the Development of CGA

To understand DHH students' literacy skills and to support their development, teachers and language therapists need validated assessments to understand the baseline performance of their students before effective interventions can be ensured (Cannon et al., 2011). An objective grammatical assessment in written Chinese can also help identify the strengths and weaknesses of individual DHH students, especially when the sociolinguistic context in Hong Kong is unique. While most children speak Cantonese, the written language they learn at school is the written form of Mandarin Chinese, which follows a very different grammatical system from Cantonese (Wang, 2019). For example, as a dative construction that encodes transfer activities, not only the main verb *give* is different in Cantonese and written Chinese, but the syntactic forms in Cantonese (1) and written Chinese (2) are also different. Many sentences that are written in Chinese but following Cantonese word order are ungrammatical (see sentence (3) as an example).

(1) 我 畀 本 書 佢 (Cantonese)
I give CL book him
'I give him a book.'

(2) 我 給 他 一 本 書 (written Chinese)
I give him one CL book
'I give him a book.'

(3) *我 給 一 本 書 他 (in Cantonese word order)
I give one CL book him

Children in Hong Kong may experience additional challenges when they are to develop their literacy following the grammar of Mandarin Chinese. Though most typically developing (TD) children are able to distinguish between the two grammatical systems when they have sufficient input through extensive reading. However, when children are growing up with insufficient guidance and input, they may still experience difficulties getting through this transition period from oracy to literacy. Children with restrictive accessibility to spoken language inputs like DHH children, they need to acquire Cantonese grammar through defective auditory perception, at the same time, they also need to master the grammatical system of written Chinese and gradually understand the differences between them. There is still a lack of empirical evidence to investigate how DHH children manage this development process. According to some minimal evidence, the academic development of DHH children in Hong Kong is not satisfactory and Chinese Language is still a difficult subject to them. Though the academic standards of different schools are different and may not be comparable with each other, the figure reported by The Hong Kong Society for the Deaf (2009) indicated that 41.7% of primary school children (from Primary One to Primary Six) with different degrees of hearing loss failed their Chinese Language examination. The result still reflects an alarming phenomenon regarding the academic and literacy development of DHH children. According to the qualitative comments from the interviewees in this study, teachers commented that DHH students' problems with speech perception and communication seemed to be the most prominent factor affecting their literacy development in school.

1.5 Aim of this Study

Before we can provide effective interventions to support DHH children's literacy development, we need to understand their genuine needs. Current assessments in written Chinese are mainly based on pre-set curriculum from an educational or functional perspective. School examinations or tests are required to cover all different elements prescribed in the curriculum. Grammatical knowledge in Chinese may basically be assessed based on some specific items such as the use of questions or logical connectives like *因爲* 'because' or *所以* 'so'. However, the coverage of specific sentence structures or grammatical knowledge is limited. There is no comprehensive assessment that can help teachers understand how well their students have acquired basic Chinese grammar especially their morphosyntactic properties. Examination results is not a good indicator to determine their grammatical development and suggest grammar-based interventions for the students.

It is a challengefor educators or clinicians to support DHH students going through the transition from oracy to literacy, especially when they are still struggling with their first language. To date, clinical efforts generally go to early identification and language interventions. Little has been done to develop scientific measurements to document DHH students' grammatical development in written Chinese. There are a few oral language assessments that comprise of items assessing students' morphosyntatic knowledge of Cantonese, for example, the Hong Kong Cantonese Oral Language Assessment Scale (HKCOLAS) (T'sou et al., 2006) or the Hong Kong Test of Preschool Oral Language (Cantonese) (TOPOL) (Wong et al., 2019). So far, no assessment tool in Hong Kong is available for measuring Cantonese-speaking children's grammatical knowledge in written Chinese, for both DHH and typically developing (TD) children.

This study is an extension of a research project "Profiling Chinese Grammatical Knowledge of Deaf and Hard-of-Hearing Students in HK and China - A Comparative Study" to compare and examine the acquisition of grammatical knowledge of Mandarin Chinese by DHH children from Hong Kong and China (Tang et al., 2020). It aims at developing and validating an assessment tool, namely the Chinese Grammatical Assessment (CGA), for measuring the grammatical knowledge of Cantonese-speaking children in written Chinese. Based on the norms developed from assessment data of TD students at local primary schools, the study also evaluates how the assessment is valid and reliable in testing DHH students' grammatical knowledge and linguistic properties in written Chinese.

Based on the above background, we would like to address the following questions in this study:

i) Is the Chinese Grammatical Assessment (CGA) valid and reliable for measuring Chinese grammatical knowledge of Cantonese-speaking primary school children in Hong Kong?

ii) Are the two CGA short tests comparable and reliable for assessing TD and DHH students' grammatical knowledge in written Chinese?

iii) Are the norms set up for CGA effective in identifying DHH students who are in need of immediate support for Chinese grammatical development?

iv) Can the results of CGA be a significant predictor of DHH students' academic performance in Chinese Language which is a major subject in primary education in Hong Kong?

1.6 Significance of the Development of CGA

With limited access to oral Cantonese, many deaf or hard-of-hearing children face severe language delay (Lau et al., 2019). These language delays often turn into deficits in print literacy after they enter the schools (Cannon et al., 2016), especially when DHH students are always delicate in acquiring a complete oral language system by the end of the critical or sensitive period of language development (Berent, 2004). When written Chinese is considered a different or second language system to DHH children, their ineffective mastery of Cantonese may also be a factor affecting their development of written Chinese.

As mentioned before, DHH children experience additional difficulties in comprehending and producing complex grammatical structures in Cantonese such as passives and relative clauses (Yiu, 2012; Lam, 2015; Lau et al., 2019). The results are similar to the studies in other languages, for example, English and Italian (see de Villiers, de Villiers, & Hoban, 1994; Blamey et al., 2001; Friedmann & Szterman, 2006, 2011; Friedmann et al., 2008; Volpato, 2010 and among others). Even though the stimuli were presented in written form, DHH students' comprehension and production of grammatical structures are still problematic (see Quigley et al., 1976; Wilbur, Goodhart, & Montandon, 1983; Berent, 1988, 1996; Lillo-Martin, 1998; Mann, 2007; Berent & Kelly, 2008; Takashi et al., 2017; and among others).

The field of deaf education continues to struggle with the development of effective instructions that can enhance DHH students' knowledge of written grammar (Cannon et al., 2016). As mentioned above, this study aims to develop the Chinese Grammatical Assessment (CGA) as the first assessment tool of its kind in Hong Kong to help assess DHH students' grammatical knowledge or morphosyntactic development in written Chinese. The norm of CGA was developed based on data of over 900 typically developing children from nine local primary schools.

At this stage, this assessment focuses on students' receptive grammar first. No language production is required during the whole process. With the assessment, local schools or rehabilitation services centres that are supporting DHH children can have better understanding about DHH students' ability to comprehend different grammatical structures in written Chinese, irrespective of their auditory ability and oral language skills. More importantly is that teachers and speech and language therapists can use the assessment to track the progress of individual DHH students and provide them immediate interventions whenever necessary (Canon & Hubley, 2014).

With reference to the Comprehension of Written Grammar (CWG) Test developed by Easterbrooks (2010), the development of a well-validated grammatical assessment is essential for:

> (a) planning whole class and differentiated instruction activities, (b) gathering information for a language report to accompany an Individualized Education Plan, and (c) charting student progress on specific grammar structures throughout the school year. (Cannon & Hubley, 2014, p.6)

It is the objective of this study to explore how CGA is effective in assessing and identifying the needs of DHH students in their Chinese grammatical development. After further validation of the assessment and research development, CGA can also be used to document the development for students with special educational needs other than deafness, for example, dyslexia, intellectual disability, or autism spectrum disorder (ASD). The assessment may also be used for students in other places, in which Chinese is an official written language such as mainland China, Macau, and Taiwan. Of course, psychometric validation according to local data from different places and developing separate norms for respective populations are essential.

Another objective of this study is to investigate how students' Chinese grammatical knowledge may impact on their academic performance in Chinese Language. If the relationship between

CGA and academic performance is positive, the results can also provide insights for educators and clinicians to identify students' needs of academic support through a quick test on their Chinese grammatical knowledge. In this regard, the establishment of local norms for CGA is not only for students' grammatical development but also academic support in Chinese Language.

Chapter 2: Literature Review

2.1 Language Ability and Reading Development

According to Chomsky, language is biologically based. Every human child is born with a language acquisition device (LAD) that readily processes auditory input necessary for the development of speech and language (Chomsky, 1957). Children acquiring their first language are based on natural inputs from the environment, which triggers the language module of the brain and sets up the principles and parameters automatically for the target language. In other words, language acquisition is based on positive evidence accessible to children. The process of language acquisition is effortless, and children are able to achieve uniform success in ultimate attainment within a short span of time (Guasti, 2002). The linguistics knowledge acquired by children forms an essential part of their grammatical system and becomes the significant foundation supporting their literacy development. Early reading is the first step toward literacy.

Reading is a complicated cognitive process requiring a simultaneous top-down and bottom-up process (Barr, Sadow, and Blachowicz, 2002). Grammatical knowledge, together with other decoding skills and some "top-down" knowledge during this interactive process, plays an important role to support linguistic comprehension of the written texts. It serves as a "collection of lexical and syntactic features" that determines accurate language comprehension and production (Cai, 2014). As summarized by Kelly (1996), reading comprehension is a combined cognitive process including:

> (1) predictable combinations of letters, (2) letter-sound correspondences, (3) the inter-word relations specified by sentence syntax, (4) the word meanings in a reader's vocabulary, (5) sentence semantics, (6) the discourse structure of stories or expository

> materials, (7) knowledge recently acquired from reading earlier parts of a text, and (8) domain or world knowledge acquired through prior reading or experience (p.75).

Research suggests that grammatical knowledge is a crucial component of a language that predicts children's language development and reading abilities (Takashi et al., 2017, p.88). "If a reader has limitations in applying the coalescing function of syntax, then phrases of text often must be maintained as strings of discrete word-units, increasing the storage burden on working memory, leaving less capacity for other processes" (Kelly, 1996, p.87). Failure to acquire grammatical knowledge severely deteriorates children's linguistic comprehension as well as their decoding strategies (Hoover and Gough, 1990). The impact is doubled.

2.1.1 Phonological Awareness versus Grammatical Knowledge

Many studies have attempted to unravel which component(s) of language knowledge predicts literacy development especially reading, but so far, no simple conclusion can be made. Some studies highlighted the significance of phonological awareness on DHH children's literacy development (Harris & Beech, 1998; Easterbrooks et al., 2008; James et al., 2008). Bus and van Ijzendoorn (1999) suggested in a meta-analysis that phonological awareness training facilitates TD children's early reading development, explaining only 12% of the total variance of word identification skills, and the impact drops to less than one percent in a long run.

As reported in another meta-analysis conducted by Mayberry, del Giudice, and Lieberman (2011), reviewing the results of 57 studies conducted in different countries (with a total of 2078 deaf participants aged 4 to 62 years) found that both phonological skills and overall language ability (both vocabulary and syntactic skills) of students significantly predicted their reading ability, but the total variance explained by overall language ability (35%), no matter in sign language or spoken language, is higher than that of phonological skills (11%). The results were similar to DHH students who are using advanced hearing technology like cochlear implants, which was an electronic device inserted into children's cochlear to stimulate the residual hair cells was proved to be a device that can successfully enhance DHH children's auditory accessibilty to speech information (Lee, van Hasselt, & Tong, 2010). As indicated in Geer (2003), 26% of the total variance of DHH students' literacy development was contributed by "phonological processing", but 47% of the total variance was explained by students' overall language ability including vocabulary and syntactic abilities. In addition, overall language ability was found to play a more significant role than phonological awareness in literacy development, especially when participating students were studying at higher-grade levels (Mayberry, del Giudice, & Lieberman, 2011).

Chinese is a logographic language system. Its nature is very different from a phonetic language like English or German. The relationship between phonological coding and reading ability in a phonetic language like English is very different from that in a logographic language like Chinese (Ku & Anderson, 2003; Tong et al., 2009; Ching & Nunes, 2015) because of their different forms of writing systems. Even for the various Chinese societies, the language learning environments are very different across different places. Even when a phonological system was adopted, the coding systems they used very different. Mainland China and Singapore adopt "Pinyin" as their phonological coding system in language teaching. In Taiwan, another system called Zhuyin Fuhao is used (McBride-Chang et al., 2012). In contrast, no specific coding system in Hong Kong is used to support children's character learning or recognition skills in Chinese. Whenever children are introduced to a new Chinese character, the only strategy children can use is to simply recognize and memorize them, using the principle of “look and say”, and relate the unanalyzed visual images of the characters to their pronunciation (McBride-Chang et al., 2012, p.95).

No particular coding system used in the Hong Kong education system does not mean that phonological awareness has no significance in language education or development in local Chinese Language education. Research studies demonstrated that phonological awareness, or access to a language’s phonological or sound system is associated with Chinese character recognition (Chow, McBride-Chang, & Burgess, 2005). The significance of “lexical tone” (one specific aspect of the phonological system in Cantonese or Mandarin Chinese) in word reading is also supported by different studies in mainland China (Shu, Peng, & McBride-Chang, 2008) or in Hong Kong (McBride-Chang et al., 2008).

So far, there are limited research focusing on the impact of phonological skills on reading ability of deaf or hard-of-hearing populations in Hong Kong. In Cheung, Leung, and McPherson (2013), 34 DHH students were given different language tasks regarding their auditory discrimination, and use of phonological and orthographic codes in word reading. Results found that the auditory discrimination ability of the DHH participants accounted for 49% of the total variance of participants’ reading ability when the effects of other variables like age, nonverbal intelligence, and hearing threshold were controlled. Auditory discrimination is found to be a substantial factor affecting DHH students’ reading. When the two strategies: the use of phonological and orthographic codes were compared, low-ability DHH readers showed a preference of orthographic coding over phonological coding (Cheung, Leung, & McPherson, 2013); even though both phonological and orthographic coding are both important information for word reading, Chinese-speaking DHH students seemed to have a preference of lexical orthographic over phonological coding in word processing.

Earlier studies confirmed that morphological awareness is an essential language component for Chinese reading comprehension. Enhanced morphological awareness substantially improves children’s literacy in Chinese (Wu et al., 2009), though phonological awareness also contributes to successful Chinese reading (Taylor, 2002). Morphological awareness was found to have a unique role in word reading, for example, compounding is characterized in Chinese word formation (Pan et al., 2021). Further research is required to investigate the relationships between morphological awareness and reading development of Chinese-speaking DHH students.

2.1.2 Vocabulary Knowledge versus Grammatical Knowledge

According to the Simple View of Reading (SVR), reading is an interaction between word decoding and recognition as well as linguistic comprehension. The former plays a more significant role in the early stage of reading development, whereas the latter contributes more to the later stages of reading development (Chan & Yang, 2018). The significant role of vocabulary knowledge and syntactic skills in Chinese reading are both confirmed in the typically developing population (Chik et al., 2012; Zhang et al., 2012) though some studies found vocabulary knowledge a strong predictor of reading (Alderson & Kremmel, 2003).

Vocabulary knowledge is a crucial factor affecting DHH children’s reading performance (Brisbois, 1995; Yamashita, 1999; Qian, 2002). It also interacts significantly with DHH children’s morphosyntactic or grammatical knowledge when their performance in reading comprehension is concerned (Kelly, 1996; Gaustad & Kelly, 2004). With the participation of 25 Chinese-speaking second-grade DHH children, Chan and Yang (2018) found that the degree of hearing loss is significantly associated with DHH students' reading comprehension. In their study, receptive vocabulary knowledge plays a more crucial role than linguistic comprehension, which comprises of both receptive vocabulary knowledge and listening comprehension (Chan and Yang, 2018). Further investigation found that receptive vocabulary knowledge contributed to early reading comprehension more than listening comprehension.

Vocabulary knowledge is nonetheless, one of the major cognitive processes involved in reading (Kelly, 1996). However, the role of grammatical knowledge in children's reading development may still be underestimated. Kelly (1996) conducted a comprehensive study with a large group of DHH adolescents from different educational settings, including those from oral school programmes (100 adolescents), total communication programmes (113 adolescents) and a postsecondary institution using total communication (211 adolescents). Kelly (1996) found that the interaction between "syntax" and "vocabulary" is the strongest predictor of reading comprehension, as compared to the two predictors alone. The results further acknowledge that both vocabulary knowledge and syntactic knowledge are important components affecting reading development of DHH adolescents. More importantly, their significant interactional effects re-iterated that grammatical knowledge and word knowledge should be given the same weight when intervention programmes are prepared for DHH children. In addition, according to Kelly's (1996) study, special attention should be given to the finding that when students' insufficient syntactic knowledge would significantly suppress the contribution of students' vocabulary knowledge to DHH students' reading comprehension. Therefore, good vocabulary knowledge alone is a factor supporting reading development. Syntactic knowledge has to be developed hand-in-hand with vocabularies.

More challenging is that DHH children's grammatical knowledge or morphosyntactic knowledge is, in general, highly vulnerable, especially in the structures that involve functional categories, the relatively "unstressed" components of the grammar (see Quigley et al., 1976; Wilbur, Goodhart, & Montandon, 1983; Berent 1988, 1996; de Villiers, de Villiers, & Hoban, 1994; Lillo-Martin, 1998; Friedmann & Szterman, 2006, 2011; Volpato, 2010; Guasti et al., 2014; Yiu, 2004, 2012; Lam, 2015, and among others). As reflected by Duchesne's (2016) review of a set of studies on the development of grammatical competence in DHH children who have received the cochlear implantation before age 2, many children with good vocabulary knowledge still faced great difficulties in their development of grammatical competence, and the struggle could be extended to students' adolescence.

Poor oral language skills are clearly predictive of poor literacy skills (Moeller et al., 2007; Lederberg, Schick, & Spencer, 2013). "Inaccurate syntactic knowledge and vocabulary knowledge have been documented as exerting a direct and adverse effect on the comprehension of many deaf readers" (Kelly, 1996, p.78). When a reader has limited knowledge in different grammatical functions of a language, phrases and sentences are only strings of discrete word units whose meaning is either vague or inaccurate. DHH students with problems of functional categories often appear to use shallow processing in their reading, that is, to extract the meaning of the sentences based on lexical categories like nouns, verbs or adjectives without a thorough understanding of the grammatical relations projected by the language's morphosyntax (Cannon et al., 2016). For example, when a deaf child responds to a sentence like *Give daddy an orange or an apple*, the child may act it out by giving daddy both an orange and an apple without noticing the functions and meanings of the logical disjunctive *or* in the sentence. They may make an interpretation based on surface word order with no attention to the semantic implications of the function word *or*. As discussed earlier, DHH students' difficulty in comprehending Cantonese grammar was also well-noted in Lau et al.'s (2019) study.

Grammatical knowledge is considered an essential building block of DHH children's reading development (Kelly 1996), however, there are limited studies investigating the grammatical development of DHH students in written Chinese. In the section below, some specific problems with morphosyntax facing people with hearing loss will be discussed in more details.

2.1.3 Impact of Deafness on Language Development

Around 95% of deaf and hard-of-hearing children are born to hearing parents. Natural oral language input should be readily available for them. However, for different reasons such as misconceptions about sign language, physiological constraints in the auditory system, delayed diagnosis of hearing impairment, or ineffective hearing aids, etc., DHH children often experience difficulties accessing to enriched spoken or signed language input during their early ages (de Villiers, de Villers & Hoban, 1994; Humphries, et al., 2012). Deafness, as a significant blockage of auditory and speech inputs, has debilitating effects on children's language development when they are developing the mental grammar of a target language (TL) during the critical period of language development. Ineffective exposure to language input can be a reason that causes the delayed development of DHH children's grammatical knowledge (Friedmann, Szterman & Haddad-Hanna, 2009). However, the major problem may not be simply a matter of exposure, but an issue of "(in)accessibility" to positive language input in their daily live environment (Berent, 2004). Children's limited hearing and speech perception abilities, delayed diagnosis and interventions, lack of sign language input from deaf adults may all be prominent reasons for the problem.

Eric Lenneberg in the 60s pioneered the study of atypical populations including children with focal brain damage, mental retardation, and deafness to investigate the nature of language acquisition (Tager-Flusberg, 1994). As discussed in the paper, a fundamental question that linguistics studies were asking is to see whether the language of atypical population would be developed in a similar or a truly "deviant" pathway (Yiu, 2015). There is an assumption that "[i]f language development looks very similar across groups of children….then this suggests that there are some fundamental constraints on the process of language acquisition that are independent of broader cognitive or social developments" (Tager-Flusberg, 1994, p. 4). Lillo-Martin (1992), with reference to Hyams (1987), Pinker (1984) and Lebeaux (1987), suggests that every child's grammar is a possible adult grammar. Instead of having a qualitative different language system (Radford, 1990), the same in-built universal grammatical system should be the same. In this regard, the physiological barriers to accessing oral language input delays DHH children's language acquisition, especially in some complex morphosyntactic properties (Lillo-Martin, 1992). From a bilingual perspective, with a full-fledged development of sign language as an accessible first language, deaf or hard-of-hearing children can still develop linguistic competence in any written language with their knowledge in sign language (Humphries, et al., 2012). Some deaf children with early cochlear implantation, can develop comparable reading comprehension skills and it benefits more if the implants are completed earlier (López-Higes, Gallego, Martín-Aragoneses, & Melle, 2015) though a very diverse result is still observed (Geers et al., 2009).

2.2 Factors affecting language acquisition of DHH children

"[T]he variables influencing deaf children's early language input and their language development are numerous and complex" (Berent, 1996, p.470). DHH children is a heterogeneous group with prominent individual differences. Deafness interacts with other developmentally and psychosocially significant variables that affect the performance of different individuals (Jamieson, 1994).

In general, the more severe the hearing loss, the more debilitating the impacts on DHH children's language development (Friedmann & Szterman, 2006). Lau et al.'s (2019) regression analysis found that the degree of hearing loss significantly predicted DHH children's oral language abilities in Cantonese. Blamey et al. (2001) found that degree of hearing loss only correlated to the speech perception of DHH children instead of their overall language ability. In contrast, Lee et al. (2010) suggested that speech perception is inevitably a good predictor of oral language ability. DHH children in an integrated education setting, in general, have a better oral language ability than those in the special school setting (Clarke, Rogers, & Todd, 1981), but the duration of mainstreaming did not associate with better oral language skills (Lau et al., 2019), possibly because of their ineffective social integration in the regular schools.

According to Geers (2004) and Nikolopoulos et al. (2004), advanced technology like early cochlear implantation successfully raised the language scores of DHH children, and the earlier the children were implanted, the better the performance they had (López-Higes et al., 2015). However, this does not mean that advanced hearing technology can fully fix or restore students' hearing loss. Many children with early implantation still lag behind their hearing peers in both of their receptive and expressive language (Geers et al., 2009), especially when their morphosyntactic development was concerned (Hay-McCutcheon et al., 2008).

Besides factors like the degree of hearing loss, educational settings, and choice of hearing device, factors related to family backgrounds such as maternal education and family involvement (Ching et al., 2013; Watkin et al., 2007), parents' sensitivity to children's communication needs and the quality of parent-child interactions (Marschark, 1993) also significantly affect their DHH children's language outcomes.

2.3 Grammatical Knowledge of the Deaf

Morphosyntax of a language, which combines syntax (rules for forming sentences) and morphology (rules for forming vocabulary), plays a vital role in a learner's linguistic comprehension (Cannon et al. 2020, p.127). Studies on the morphosyntactic development of deaf children have emerged since the 1970s. Quigley et al. (1976) used the Test of Syntactic Ability (TSA; Quigley et al., 1978) to assess a group of 450 profoundly deaf learners, aged from 10 to 18 years old, in their comprehension and production of 22 different English syntactic structures in written form. They found that their morphosyntactic development of English was considerably delayed in all 22 structures, and their performance was even inferior to that of the 60 younger normal-hearing children aged 8-10 years old. Table 1 is a summary of their findings grouped under nine major grammatical categories (Quigley & Paul, 1994, p.164-165).

The comprehensive study by Quigley et al. (1976) brought deaf educators' attention to the tremendous difficulty DHH learners have in their acquisition of English morphosyntax. Even though Quigley et al. (1976) found differences in DHH students' performance between the Age 10 and the Age 18 groups in all nine grammatical categories (see Table 1), the average scores of the Age 18 group in many syntactic categories like "Disjunction and Alternation", "Complementation" and "Relativization" was lower than that of the younger hearing group aged 8-10. Years later, Wilbur, Goodhart and Montandon (1983) tested nine more structures different from the 22 structures included in Quigley et al. (1976). They found “Ellipsis”, “Reciprocal Pronouns” and “Comparatives” were also different grammatical categories for DHH learners. These results brought out an important observation that besides the general acquisition milestones, DHH learners seem to have additional difficulties handling some areas of grammatical knowledge.

Table 1. Performance of Students in Different Syntactic Categories in English (Quigley & Paul, 1994)

Syntactic Categories	Deaf Students				Hearing Students
	Average across ages from 10-18 yrs (%)	Age 10 (%)	Age 18 (%)	Increase* (%)	Average across ages from 8-10 yrs (%)
Negation	76	57	83	26	90
Conjunction	73	57	86	29	92
Question Formation	66	46	78	32	98
Pronominalization	60	39	78	39	90
Verbs	58	53	71	18	79
Complementation	55	50	63	13	88
Relativization	54	46	63	18	82
Disjunction and Alternation	36	22	59	37	84

* Represent the percentage increase between the Age 10 group and the Age 18 group.

2.4 Additional Challenges of Functional Categories

de Villiers, de Villiers and Hoban (1994) proposed that the central problem of the morphosyntactic development of DHH children was their acquisition of structures involving functional categories such as determiners (e.g. *this* and *that*), inflectional morphemes (e.g. -*s*, -*ed* and -*ing*) and complementizers (e.g. *wh*-words) (Radford, 2004). The functional categories characterized by their phonologically “unstressed” nature create more difficulties for DHH learners in identifying them audiologically. Unlike lexical categories, nouns, verbs, and adjectives, functional categories are “closed-class” items, possessing little or no semantic information, but play a vital role in a language that provides a skeletal structure to accommodate the content words (Lust, 2006). They “serve primarily to carry information about the grammatical function of particular types of expression within the sentence” (Radford, 2004, p.40), or in other words, they “organize grammatical relations between words within a sentence and between sentences within a text” (Takashi et al., 2017, p.91). For example, the possessive 嘅 (*ge3*) in Cantonese or 的 (*dik1*) in written Chinese does not represent a specific object or an action, but they can bring out the positive relationship between the possessor and the objects he or she possessed. More examples in Chinese will be given below to illustrate the concept.

Functional categories in a language are, in fact, incorporated in a wide range of grammatical structures as an essential component of a language. As the grammatical structures of Chinese are different from English, the errors made by DHH children in English may not appear in the same ways as in Chinese. The following examples are provided to help illustrate some major linguistic properties in written Chinese.

2.4.1 Acquisition of the Argument Structure

Based on Quigley (1969), the analysis of written samples from English-speaking deaf participants show that they have difficulties in: (i) the use of auxiliary verbs, (ii) the use of tense markers, (iii) the use of copulas, and (iv) the obligatory nature of verbs. Each of these areas will be briefly discussed with some examples below in Chinese.

i) Auxiliary

In a grammaticality judgment test conducted by Quigley, Montanelli and Wilbur (1976), only

45% of the Age 10 profoundly deaf children correctly pointed out that sentences with a missing auxiliary verb were ungrammatical. By conducting an elicited production task for wh-questions, de Villiers (1988; cited in de Villiers, de Villiers, & Hoban, 1994) found that deaf children aged 6-14 produced more syntactic "errors" than normal hearing children aged 3-5 even though they were much younger than the deaf children. The most prominent error (82.4% of all the syntactic errors) of the deaf participants was the omission of auxiliary verbs such as *is, am, are* in a sentence. Deaf children seemed unable to consider auxiliaries as a significant syntactic constituent of a sentence and so they just simply ignored their existence.

是 'be' is a copular verb. It is one of the major auxiliary verbs used in Chinese. The basic use of the auxiliary *是* 'be' in Chinese is quite similar to English, like *我是一個男孩子 (I am a boy)*. Unlike English, Cantonese *be* requires subject-verb agreement, which means the verb form has to change according to the subject and the tense of the sentence. In the following examples, *are* agrees with the plural subject *we* in the sentence *We are girls* and *was* agrees with the past tense of the sentence *She was a teacher*. No morphological changes in the verbs are required in Chinese in terms of subject agreement or tense agreement. However, this may create additional difficulties to understand the meaning behind the concept of "agreement" in English though this is not the major concern on this study.

Different languages have different forms of linguistic complexity, no matter whether morphology or syntax is concerned. Further investigation is required to explore whether Chinese DHH students in Hong Kong would face the same level of difficulties as English-speaking DHH children do.

ii) Tense Marking

Besides the problems of auxiliaries, English-speaking DHH children also face difficulties in tense marking. Morphological changes in the verb are induced based on the tense and aspect of a sentence, e.g. *He has finished reading two books*. The verb *finish* has to change to *finished* as a tense marking, representing a completed action. DHH children often fail to identify sentences with omissions of tense marking as ungrammatical (only 60% correct across all ages from 10 to 18 years) (Quigley, Montanelli, & Wilbur, 1976). Compared with the TD group (aged 4;0 to 5;6), DHH children's (aged 6;4 to 13;4) performance was inferior to the TD group. DHH children would produce less regular past-tense marking, and more errors with unmarked tense in English (Baumberger, 1986).

There is an absence of explicit tense marking in Cantonese (Matthews & Yip, 2011) and Mandarin Chinese (Li & Thomson, 1940). The problem in tense marking in Chinese may not be relevant to Cantonese-speaking DHH children. Taking the above sentence as an example, 了*(liu5)* is used as a perfective marker after the verb like *他看了兩本書* 'He has finished reading two books' to represent the time-bounded event (Li & Thompson, 1940). Therefore, the problem facing Cantonese-speaking DHH children in Hong Kong may not be the same as English-speaking children. Their problems in tense and aspect may rest on the use and understanding of the prefectural marker *了 (liu5)* as a single word to indicate the completed action.

iii) Obligatory Nature of Verbs

Below are two sentences written by an 8-year-old Cantonese-speaking deaf girl. As observed, she may simply neglect the obligatory nature of verbs in the forms of omission of the main verb like sentence (4), and duplication of the main verb like sentence (5).

(4) * 爸爸 巴士 回家

father bus back home

(5) * 媽媽 吃 食 飯

mother eat eat rice

In sentence (4), the main verb is actually missing though the noun *巴士* 'bus' may be used as a verb by the deaf child. In (5), a duplication of the main verb *eat* is observed' According to the sentence, both the word *吃* 'eat' and *食* 'eat' have the same meaning of eating. Though the examples in (4) and (5) only represent the data of an individual deaf child, research does find some similar evidence of verb omission or duplication in DHH children in other languages (de Villiers, de Villiers, & Hoban, 1994).

Quigley, Montanelli and Wilbur (1976) reported that DHH children were able to judge sentences with an omission of main verb as ungrammatical. However, when they were asked to rewrite the sentences, 33% of the sample group did not insert a verb in the sentences. The obligatory nature of verbs is not recognized by many DHH children. Besides, in a study of Hebrew speakers, Hana and Esther (1998) found that the "omission of the subject or the main verb in a sentence" was the syntactic deviation frequently noticed in DHH children's language production.

2.4.2 Acquisition of Complement Phrases (CP)

Research studies found that DHH children perform better in single-clause structures like *The boy broke the window* than more complex complement constructions like *The boy asked his mother if he could go outside* (Quigley, Montanelli & Wilbur, 1976; Hana & Esther, 1998). The phenomenon reflects DHH children's difficulties in handling complex sentences, especially their inability to comprehend or produce Complement Phases (CP) (de Villiers, de Villiers, & Hoban, 1994). The insights obtained from English-speaking DHH children help identify suitable structures that should be included in the grammatical assessment to be developed for Cantonese-speaking children.

i) *Wh*-questions

Wh-questions in English involve a syntactic process called *wh*-movement. From a linguistic perspective, it involves a syntactic movement of the *wh*-word (such as *why*, *what*, *when*, etc.) into the specifier position of the CP node. According to de Villers, de Villiers and Hoban (1994), DHH students had no great difficulty in producing the *wh*-questions with the *wh*-words situated in the initial position of the sentence, but they often (65% of the trials) omit the auxiliaries in the questions like **Where the cat?* rather than *Where is the cat?*

Chinese *wh*-questions basically involve no overt movement of the *wh*-words (Law, 1990). In addition, the auxiliary stays in situ in the original position in the *wh*-questions (see sentence (6)), which follows the basic word order of a declarative sentence like (7). The construction is similar to that of the echo-questions in English.

(6) 小 明 下 星期 離開 美國

Siu Ming next week leave America

‘Siu Ming will leave America next week.’

(7) 小 明 何時 離開 美國

Siu Ming when will leave America

‘When will Siu Ming leave America?’

The use of auxiliary in English is a major issue facing DHH children, but it seems that there is no similar evidence found in Cantonese-speaking children. Instead, according to anecdotal observation, Cantonese-speaking DHH children always find it hard to grasp the meaning of the different sentence-final particles in Cantonese such as *嘅* (*ge3*), *囉* (*lo3*) and *嚸* (*bo3*) as they are relatively unstressed auditorily. A change of the particle in Cantonese will create a change in the sentence’s meaning. Whether DHH children may have difficulty in the correct use of different question particles like *嗎* (*maa1*) and *呢* (*ne1*), further assessment and investigation is required. Will it be easier for DHH students when these words are presented in a form of written Chinese? No research finding can be consulted at present.

2.5 DHH Students’ Grammatical Development in Written Chinese

Cross-linguistic evidence has confirmed that DHH children and/or adolescents have delayed acquisition in many different morphosyntactic structures when compared with their hearing counterparts in different languages, such as English (Berent, 1996; Nikolopoulos et al., 2004), French (Tuller, & Jakubowicz, 2004), Italian (Volpato, 2010), Hebrew (Friedmann, & Szterman, 2006; Friedmann, Szterman, & Haddad-Hanna, 2009), Cantonese (Yiu, 2012) and written Chinese (Lam, 2015). Despite the differences in the morphosyntactic structures that they studied, the general picture is obvious that deafness has created additional barriers for DHH students to acquiring grammatical knowledge of their first language, which is for most children, an oral language.

According to an initial analysis of the grammatical development of by typically developing and hard-of-hearing students in written Chinese (Tang, Li, Li, & Yiu, 2023), no significant difference was found in DHH primary school students' comprehension of some structures like Negation, Passives, and Comparatives, Aspect, Locative Constructions and Modals. Do they experience additional barriers or problems in their development of grammatical knowledge especially those with functional categories and complex structures? Do TD and DHH students have similar developmental pathways in written Chinese? Further studies and data are required to better understand their similarities and differences in Chinese grammatical development.

2.6 A Preliminary Summary

A learner's overall literacy development is associated with their acquisition of functional grammar (Cannon et al., 2020). For DHH learners, even when their hearing loss is diagnosed in their early ages and they are fitted with advanced hearing aids or cochlear implants, they remain in a relatively disadvantaged position in their reading development (Chan & Yang, 2018). A more recent study revealed that the reading ability of two cohorts of DHH children with a 10-year age difference, however, no significant difference was found between their reading achievements. Both two groups of students lagged behind their normal-hearing counterparts (Harris, Terlektsi, & Kyle, 2017). Cannon et al. (2016) suggested that DHH students' lack of bottom-up skills like syntactic knowledge may be the main reason for their overall suppressed literacy development. In this regard, starting from the 1970s, researchers have developed different assessments to help investigate the major problems in morphosyntax facing DHH learners. In the following sections, we will briefly describe the language policy in Hong Kong and how it may impact on DHH students' literacy development. We will then discuss the availability of grammatical assessments for DHH children in Hong Kong and in other countries.

2.7 Language Policy in Hong Kong

Hong Kong has adopted a "biliterate and trilingual" language policy. Students are expected to master written Chinese and English and to speak Cantonese, Putonghua, and English. Regarding the sociolinguistic situation, Hong Kong has led to an unique phenomenon of language use in society. According to the *2021 Population Census* of the Hong Kong SAR Government (Census and Statistics Department, 2022), Cantonese, as a dialect of Chinese, is spoken by about 88.2% of the population aged 5 and over while only 2.3% of the population use Putinghua. Chinese printed texts adopt Cantonese pronunciation and traditional characters. Currently, there are some trends to promote using Putonghua to teach the subject of Chinese Language in school. Nevertheless, oral Cantonese remains the prominent medium of instruction in the Hong Kong education system.

Hong Kong practices universal neonatal screening and early identification. DHH children are prescribed with hearing aids by the Education Bureau. Some children with severe to profound hearing loss receive cochlear implantation with the recommendation of the Ear, Nose, and Throat (ENT) specialists of the Hospital Authority. Cochlear implantation is an electronic device that is inserted into the cochlea through surgery to stimulate the hearing nerves that transmit the sound signals directly to the brain. For children with specific problems with their cochlear and/or auditory nerves, such as the absence of cochlea or auditory nerve, cochlear implantation may not enhance their speech perception. For the last ten or more years, an auditory brainstem implant which requires a surgery to put the electrodes onto the brainstem, is now also considered an alternative option for deaf children (Colletti & Shannon, 2005).

All DHH children are referred for speech and language training in oral Cantonese, with a waiting time of about 6 to 9 months after diagnosis. When they reach age 3 or 4 during preschool education, they are taught how to read written Chinese words or sentences using Cantonese pronunciation. Literacy training is more intensive when children enter primary education at around age 6. DHH children are expected to learn written Chinese and its grammar based on their knowledge of oral Cantonese. However, this transition poses an additional challenge to DHH children as they are required to cope with the development of oracy and literacy skills grounded in two different linguistic systems (Lau et al., 2019). In the study *A Survey on the Difficulties and Challenges Encountered by Primary Students with Hearing Impairment in Integrated Education* (The Hong Kong Society for the Deaf, 2009), it is reported that 41.7% of the children failed their Chinese examinations in their schools, and 31% of them failed in all three basic subjects including Chinese, English and Mathematics based on the schools' replies on the academic performance of 127 DHH children in the mainstream primary schools. This reflects that both DHH children's Chinese literacy and their overall academic performance do not reach the general standard or expectation of the schools. According to the teachers, communication barrier is one of the major reasons for this alarming result (The Hong Kong Society for the Deaf, 2009).

2.8 Education for the Deaf

Deaf education in Hong Kong was mainly conducted in special education settings from the 1930s to the 1970s and sign language was the major mode of communication during that period of time (Sze et al., 2012). After the White Paper “Integrating the Disabled into the Community” was published in 1977 (Hong Kong Government, 1977), DHH children were encouraged to study in mainstream schools as far as possible. According to the figures provided by the Education Bureau from 2012 to 2017, around 650-690 DHH students were studying in regular public schools (Audit Commission, 2018). Less than 10% of the DHH students are placed in the remaining special school for the deaf in Hong Kong (Yiu, Tang, & Ho, 2019).

In terms of deaf education policy, the government adopts an oral approach to education for the deaf. Sign language is not encouraged in terms of the policy, no matter in the mainstream or special school settings. Whether sign language should be used as a mode of communication or medium of instruction in deaf education practices has been a vigorous debate for decades among deaf educators in different areas of the world. Following the resolution passed in 1880 in the second International Congress on the Education of the Deaf (hereafter “ICED”) in Milan that led to the removal of sign language in deaf education around the world (Moores, 2010), the impact was also significantly affected the language policy for deaf children in Hong Kong. Under such a trend, sign language is sometimes considered a “taboo” in deaf education. It is not encouraged though it is not totally banned in Hong Kong. Oralism a dominant approach to deaf education in Hong Kong for centuries. All deaf and hard-of-hearing students undergo their early training only through listening and speaking, no matter how much perceptual limitation they are experiencing. In contrast, under an oral education philosophy, sign language is often taken as the last resort, should only be given to “failure” cases.

Cochlear implantation has significantly improved deaf children's auditory and speech performance, but its etiology is still inconsistent among individuals (Humphries et al., 2012). For different reasons, the impact of deafness continues to create barriers to communication and language development of DHH children in Hong Kong and other countries, eventually leading to cognitive delay and academic failure (see The Hong Kong Society for the Deaf, 2009; Figueras, Edwards, & Langdon, 2008; Pisoni et al., 2008; Castellanos, Pisoni, Kronenberger, & Beer, 2016).

With the enactment of the United Nations Conventions on the Rights of Persons with Disabilities (2016), the availability of sign language in deaf education is highlighted in the Article 24 on Education. Some evidence sees the positive impacts of including both signed and spoken language in support of literacy development of DHH children, with or without cochlear implants (e.g., Hermans etal., 2008; Lange et al., 2013; Rinaldi & Caselli, 2014).

There is no single approach that can guarantee success in deaf education (Marschark et al., 2015). No matter which mode of communication is adopted in the classrooms for DHH children as discussed above, whether the education processes can effectively support DHH children's literacy development is of educators' and researchers' major concern (World Federation of the Deaf, 2016). "Challenging the fourth-grade ceiling" of deaf college graduates' reading achievement is still a mission of many deaf education programmes worldwide (Mayer, Trezek, & Hancock, 2021).

Reading development plays a vital role in deaf children's education. If grammatical knowledge is one of the key factors predicting reading comprehension of DHH children (Kelly, 1996), developing an assessment that helps to keep track of Cantonese-speaking children's of grammatical development in written Chinese is of unique significance in Hong Kong.

2.9 Grammatical Assessments for DHH Learners

DHH students' grammatical knowledge is a significant factor affecting their development of literacy skills. However, quite a few assessment tools have been developed to measure morphosyntactic knowledge of DHH students, no matter using a written or auditory-oral mode of assessment. The Test of Syntactic Abilities (TSA; Quigley, Steinkamp, Power, & Jones, 1978) is one of the first few assessment tools developed to identify the strengths and weaknesses of DHH students' morphosyntactic knowledge in English. TSA focuses on nine major grammatical categories (with 20 sub-categories) including negation, conjunction, determiners, question formation, verb processes, pronominalization, relativization, complementation, and nominalization through the tasks of sentence completion, sentence correction and free writing (Quigley, 1977; Quigley, Steinkamp, Power, & Jones, 1978). Results based on TSA indicated that DHH learners consistently lagged behind their normal-hearing counterparts with a similar order of difficulty (Quigley, Steinkamp, Power, & Jones, 1978).

Following TSA (Quigley, Steinkamp, Power, and Jones, 1978), Wilbur, Goodhart, and Montandon (1983) developed another grammatical assessment, which contains nine morphosyntactic categories in 125 items, covering grammatical domains either not tested (e.g., ellipsis, indefinite pronouns) or not detailed enough (e.g., *why*-questions and modals) in TSA. Results showed that *wh*-questions were the easiest, reciprocal pronouns and ellipsis were the most difficult structures for DHH learners. DHH students were found to perceive grammatical elements like indefinite pronouns, quantifiers, modals and comparatives as superficial lexical items without in-depth syntactic analysis. All these results bring insights to deaf educators on how literacy intervention should be developed for DHH learners.

TSA is presented in a written mode to ensure no perceptual barriers to communication during the assessment process. With the development of advanced hearing technologies as well as early speech and auditory training, there are also assessment tools that measure DHH students' English grammar via the auditory-oral mode of presentation such as the Rhode Island Test of Language Structure (RITLS; Engen & Engen, 1983) and the Grammatical Analysis of Elicited Language (GAEL; Moog & Geers, 1980, 1985). A more recent assessment for measuring grammatical knowledge of implanted children is the Diagnostic Evaluation of Language Variation – Norm Referenced (DELV-NR; Seymour, Roeper, & de Villiers, 2005). DELV-NR covers primarily 26 grammatical structures. Its design of stimuli is based on theories of linguistics and language acquisition.

The Comprehension of Written Grammar (CWG; Easterbrooks, 2010), is the most recently developed assessment tool for DHH children from aged 7-11 years old. Twenty-six grammatical structures are included in CWG, covering a wide range of functional grammar that consistently challenges DHH learners. There are different reports discussing how CWG's content validity (see Cannon & Hubley, 2014), as well as its reliability and known-groups validity (see Cannon et al., 2016) for both TD learners with normal hearing ability and DHH learners. Similar to what the CGA development would do, the assessment results of hearing participants in CWG were used to set as an age-equivalent norm to track the development of DHH learners.

Assessment tools that adopt the auditory-oral mode of presentation generally aim to examine how efficiently DHH children perceive speech stimuli with prescribed hearing aids and cochlear implants. However, presenting the test items orally and soliciting an oral response from DHH subjects may take the risk of biased results due to their limitations in speech perception and speech production. In other words, a test design that adopts comprehension via a written mode of presentation has the advantage of reducing barriers arising from DHH students' auditory perception. This helps assess more precisely on DHH students' knowledge in different morphosyntactic properties or grammatical structures (Mayberry et al., 2011).

2.10 The Development of CGA

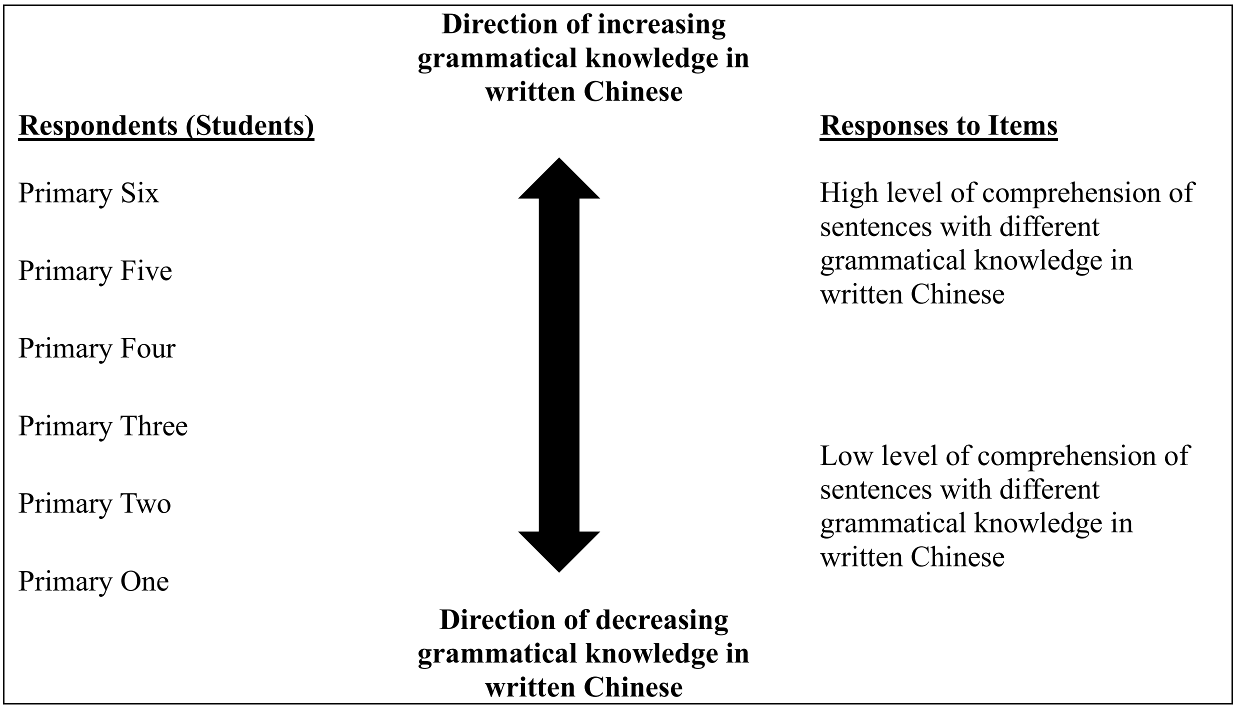
The need for developing a standardized assessment to test for deaf or hard-of-hearing (DHH) and typically developing (TD) students' Chinese grammatical knowledge in Hong Kong is well-justified. The following question is "How a valid and reliable test can be developed?" With reference to Wilson's (2005) "Four Building Blocks" model for test development, the development of two normative CGA short tests, in this study, emphasizes the significance of construct identification, the validity, reliability and fairness of the measurement based on an item response modelling approach (Efeotor, 2014). The Four Building Blocks model include i) The Construct; ii) The Item Design; iii) The Outcome Space; and iv) The Measurement Model (Wilson, 2005).

As explained by Wilson (2005), "[a] construct could be part of a theoretical model of a person's cognition – such as their understanding of a certain set of concepts or their attitude toward something – or it could be some other psychological variable such as 'need for achievement' or a personality variable such as bipolar diagnosis" (p.21). The focal concept or the latent trait that CGA intends to measure is the grammatical knowledge in written Chinese in both TD and DHH students, according to the specific context in Hong Kong.

2.10.1 The Construct

As there is no clear evidence showing how and in what developmental sequence Cantonese-speaking children acquire grammatical knowledge in written Chinese, the initial construct for the study can only be a simple structure unlike the one established by The Common European Framework (CEFR) in 2001, in which six levels of language proficiency are clearly defined with well-structured descriptors (Efeotor, 2014). No specific milestones regarding Cantonese-speaking children's development of different grammatical structures in written Chinese would be defined at this stage. The current development of CGA is indeed a first step to exploring how the construct can be further structured based on empirical data.

As a general framework to develop CGA, the comprehension of the three inter-related components of grammar, namely the form, meaning and use of different grammatical structures (Larsen-Freeman, & Celce-Murcia, 2016) are considered when the items are designed for the assessment. An assumption is that with a longer time of Chinese Language education in primary schools, Cantonese-speaking students' accuracy in comprehending sentences in written Chinese should be significantly enhanced. The construct map in Figure 1 is used at this stage to represent the basic construct of CGA.



Direction of increasing grammatical knowledge in written Chinese

Respondents (Students)

Primary Six

Primary Five

Primary Four

Primary Three

Primary Two

Primary One

Responses to Items

High level of comprehension of sentences with different grammatical knowledge in written Chinese

Low level of comprehension of sentences with different grammatical knowledge in written Chinese

Direction of decreasing grammatical knowledge in written Chinese

Figure 1. Construct Map of CGA

2.10.2 Item Design

Before collecting the norm data, a series of procedures were conducted to review similar works in the literatures, develop the items, and refine them based on the results of initial trials with the consultation support from three linguistics experts (Tang et al., 2023), trying to "link the construct closely to the items – that brings the inferences as close as possible to the observations" (Wilson, 2005, p.26). The item design is to "operationalize dimensions of the construct into items that give an accurate representation of the ability of the participants" (Efeotor, 2014, p.208). As for CGA, the items are designed in a fixed-response format including four different types of tasks: i) picture selection task, ii) truth-value judgement task, iii) grammaticality judgement task and iv) multiple choice question. Though items in a fixed-response format may not allow detailed responses for an in-depth analysis like open-ended questions, it fits the purpose of grammar testing, which avoids demanding requests for children to explain their answers using difficult grammatical terminology (Efeotor, 2014). Using a fixed-response format also possesses other advantages such as allowing shorter testing time, allowing a wider scope of items included in the assessment, and reducing scoring bias of assessors.

2.10.3 Outcome Space

Outcome space is defined as "a set of categories that are well defined, finite and exhaustive, ordered, context-specific, and research-based" (Wilson, 2005, p.65). Among the four types of questions, there are no specific distractors designed for truth-value judgement and grammaticality judgement questions. All students need to determine whether the picture matches the meaning of the sentence in the truth-value judgement questions. For grammaticality judgement questions, they need to judge whether a sentences or stimulus is grammatically correct. For picture selection and multiple-choice questions, the distractors were designed with reference to different acquisition studies for both TD and DHH students such as Lam (2016) on relative clause in Chinese; Yiu (2012, 2015) on relative clause and passives in Cantonese respectively.

Common errors made by children are, in general, reasonable distractors for the items when grammatical development is measured. In CGA, no standard number of distractors was created for the items. The choices of distractors depend on the nature of the specific items and the linguistics properties that are tested. For example, two distractors are used in the questions about the categories of morpheme distinction as in general children are found to be confused about the use of the structural particles *的 (dik1), 地 (dei6)* and *得 (dak1)*. And for the items about the five *wh*-question words *什麼時候 'when'*, *怎樣* 'how', *哪裏* 'where', *誰* 'who' and *爲什麼* 'why', except the correct response, the other four *wh*-words are all included as the distractors of the questions. No matter how many distractors are used in the test items, the scoring scheme is the same for all items: “1” mark for a correct response, and “0” mark for an incorrect response. In another word, the marking scheme is dichotomous in nature. Negative marking is not considered in CGA to avoid additional pressure on participating students and unintentionally increase the tendency of “no response” from students (Efeotor, 2014).

2.10.4 The Measurement Model

There are two main approaches to measurement, namely Classical Test Theory (CTT) and Item Response Theory (IRT). Among them, the former is test-oriented and the latter is item-oriented (Efeotor, 2014). CTT deals with the observed scores (X) and its relationships with the true scores (X) and errors (E) made in the measurement. Its analysis is dependent on the examinee samples, while IRT concerns more about how an individual person’s performance relates to individual items (Hambleton & Jones, 1993).

With the advantage that the person’s ability and item statistics can be compared on the same scale, the Rasch Model (Rasch, 1960), as a special case of IRT, is used in this study to review the original 172 items of CGA as the primary focus at this stage of test development is more on the items’ reliability and validity, rather than the measurement model (Efeotor, 2014). Especially when the sample size was small (963 TD and 40 DHH data), CTT may not be an appropriate and reliable measure in this study. The use of IRT or Rasch model, as an item-based analysis can increase reliability of the validation process and assess all items’ psychometric properties individually, which provides objective evidence for item selection and thus the development of two equivalent short tests.

In fact, besides Rasch analysis, which is used for the validation of the assessment and the test items for two CGA short tests, classical test theory is also adopted for the development of the norms of the tests, based on students' raw scores. In this regard, educators and clinicians in practice can use the assessment more easily for the review of students' performance and develop respective interventions for them.

Chapter 3: Methodology

Considering the aforementioned local background and the need for a normed test for the assessment of grammatical knowledge in written Chinese, a validated assessment, namely the Chinese Grammatical Assessment (CGA) would be developed based on the original 172-item profiling tool. After completing the validation process, a case study with a group of typically developing (TD) and deaf and hard-of-hearing (DHH) students would be conducted to review further how this test is applicable in assessing students' performance in Chinese grammatical knowledge and identifying their needs based on the results.

The study aims to achieve two major objectives:

i) To develop two short versions of the Chinese Grammatical Assessment (CGA) after validation of the psychometric properties of the original long version with a total of 172 items, based on the data from typically developing (TD) students at local primary schools.

ii) To investigate if the two short tests are reliable and valid in assessing the Chinese grammatical knowledge of a group of DHH and identifying the needs of individual students.

The study is organized in terms of five phases. They are described as follows:

i) **Phase One:** Conducting Content Validation of CGA

ii) **Phase Two:** Psychometric Review of the Items

iii) **Phase Three:** Development of Two Equivalent Lists

iv) **Phase Four:** Establishing the Norms of CGA

v) **Phase Five:** Application of the Two CGA Short Tests on DHH Students

This chapter will discuss the background of the methodology adopted for the study in the following paragraphs. Some detailed descriptions regarding the methodology used in the five phases of test development will also be described and discussed in the respective chapters.

3.1 Background of CGA Development

The development of the Chinese Grammatical Assessment (CGA) is based on the data collected from the project "Profiling Chinese Grammatical Knowledge of Deaf and Hard-of-Hearing Students in HK and China – A Comparative Study" of the Department of Linguistics and Modern Languages, The Chinese University of Hong Kong (Tang, et al., 2020), with the support of the General Research Fund (Project Number: GRF#14611315). The study was approved by the Survey and Behavioural Research Ethics Committee (Reference No. SBRE-22-0053) of the Chinese University of Hong Kong, and the use of the data for this study has also been granted Approval for Exemption from Ethical Review (Ref. E2022-2023-0011) by the Human Research Ethics Committee of the Education University of Hong Kong.

The initial development of CGA aims to develop a profiling tool to review deaf and hard-of-hearing (DHH) children's grammatical knowledge in written Chinese from a linguistics perspective. It also aims to compare and examine the acquisition of written Chinese by TD and DHH students in Hong Kong, Macau and mainland China. This study is a part of the above-mentioned project, aims to develop two short versions of CGA and their norms by conducting different well-established validation procedures. The two short tests would then be used to assess a group of Cantonese-speaking children in Hong Kong as a case study to review the practical applications of CGA. In this chapter, we will first explain the background of the development of CGA and then the key features of the assessment.

3.1.1 Basic Design

The development of the Chinese Grammatical Assessment (CGA) aims to address the needs of local DHH students with auditory and language deprivation. Grammatical knowledge is one of the major concerns of DHH children that is showed to be a significant factor affecting their reading and academic development. DHH students are considered vulnerable in acquiring grammatical knowledge, especially for those involving functional categories. However, local acquisition research for the DHH population is scarce. Our understanding of their development of written Chinese is also limited.

To the author's knowledge, no available tool in Hong Kong can be used to assess Cantonese-speaking children's grammatical knowledge in written Chinese. The motivation for developing the Chinese Grammatical Assessment (CGA) for both TD and DHH students in Hong Kong was basically to fill this gap. When the norm is established based on the data from primary TD students, the test will be useful in supporting local professionals in the field of education or speech therapy to help identify the needs of DHH students and tailor intervention plans for them.

3.1.2 Item Construction

The development of an effective tool to document primary school students' Chinese grammatical knowledge was one of the major objectives of this study. The construction of the test items for CGA drew references from formal analysis of a set of representative linguistic structures in Mandarin Chinese (including but not limited to Li & Thompson, 1981; Huang, Li, & Li, 2009; Yip & Rimmington, 2004 and among others) as well as the related research findings in the first and second language acquisition of Mandarin Chinese and Cantonese. Some studies that directly involved DHH students in Hong Kong (Yiu, 2012; Lam, 2016; and among others) and mainland China (Chan & Yang, 2018; Wang & Andrews, 2021; and among others) were also reviewed. An initial long version of CGA was then developed based on different morpho-syntactic properties in written Chinese. A comprehensive review and trials were made by the research team to see if the grammatical categories covered in the assessment were having good representativeness in terms of the coverage of Chinese grammatical knowledge for local primary school students. In addition, whether the items were relevant to the targeted linguistics properties and the item design was appropriate for primary school students in Hong Kong were also discussed.

When the design of the item pool was relatively stable, three renowned experts in Chinese linguistics and language acquisition were invited to review all the available items from a linguistic perspective. After considering their comments and suggestions, some items were revised and modified. There were also some newly developed items included in the item pool, contributing to the 172-item CGA, covering 18 major grammatical categories and 48 sub-categories of Chinese grammatical knowledge was confirmed (see Table 2 for the 18 major categories, and Appendix A for all the 172 items and their respective grammatical categories and sub-categories).

Table 2. The 18 major grammatical categories and 48 sub-categories of written Chinese adopted in CGA

Category	**Grammatical Category**		**Examples**
S01	Ba-construction	把字句	小明把花瓶打破了。 ‘Siu Ming broke the vase.’
S02	Passives	被動句	花瓶被小明打破了。 ‘The vase was broken by Siu Ming.’
S03	Binding	約束句	小明的哥哥在畫他。 ‘Siu Ming's brother is painting him.’
S04	Relative clauses	關係從句	戴著帽子的男孩在踢球。 ‘The boy in a hat is playing football.’
S05	Comparatives	比較句	小明比小華高。 ‘Siu Ming is taller than Siu Fa.’
S06	Quantification	量化句	所有男孩都在畫畫。 ‘All the boys were drawing.’
S07	Double-object Construction	雙賓句	小明送給老師一束花。 ‘Siu Ming gave the teacher a bouquet of flowers.’
S08	Locative Existential	處所存在句	操場上站著一個男孩。 ‘There is a boy standing in the playground.’
S09	Control	兼語句	小明要姐姐講故事。 ‘Siu Ming asked his sister to tell a story.’
S10	Cleft Sentences	分裂句	小明是後天參加比賽的。 ‘Siu Ming will participate in a competition the day after tomorrow.’
S11	Question	疑問句	媽媽怎麼會去學校？ ‘How did mom go to school?’
S12	Morpheme Distinction	結構助詞	小明笑得很開心。 ‘Siu Ming smiled happily.’
S13	Negation	否定句	小明昨天沒有參加比賽。 ‘Siu Ming did not participate in the competition yesterday.’
S14	Preposition	介詞	小明向公園跑去。 ‘Siu Ming is running towards the park.’
S15	Localizer	方位詞	小明坐在沙發上。 ‘Siu Ming is sitting on the sofa.’
S16	Aspect	體貌詞	小明喝了一杯水。 ‘Siu Ming has drunk a glass of water.’
S17	Question words	疑問詞	小明什麼時候參加比賽？ ‘When does Siu Ming participate in the competition?’
S18	Question Particles	疑問語氣詞	小明要不要參加比賽呢？ ‘Does Siu Ming want to participate in the competition?’

3.1.3 Data Collection

A series of data collection from 2015 to 2019 were conducted from TD and DHH students, and thus a database was set up by the Centre for Sign Linguistics and Deaf Studies, The Chinese University of Hong Kong (hereafter "CSLDS) (Tang, et al., 2020). With the consent given by CSLDS, a set of data with 963 TD and 40 DHH students from nine regular primary schools was used in this study for the development of two alternate forms of CGA (see Table 3 for the numbers and grade levels of students). No students with special education needs were included in this set of data. All the schools and parents joined the study on voluntary basis.

Table 4 shows the distributions of the 963 TD students at the nine local primary schools located in different regions of Hong Kong. Among them, there were 426, 167 and 368 students from the schools in Kowloon, New Territories West and New Territories East respectively. No schools in Hong Kong Island participated in the study. The students in the study were studying at different grade levels, from Primary One (P1) to Primary Six (P6), with a distribution from 105 (10.90%) to 225 (23.36%) students. There are more students at the junior grade levels (P1-P3) than those at the senior primary grade levels (P4-P6) in the dataset.

The data of 40 DHH students were collected from 2017-2019 at two local primary schools. No longitudinal data from DHH students were used in this study. The dataset was mainly contributed from the data collected in 2019. The schools adopted the Sign Bilingualism and Co-enrolment in Deaf Education (SLCO) Programme, in which there was a bigger group or a critical mass of DHH students, relative to other mainstream schools in Hong Kong (Tang & Yiu, 2013; Yiu, Tang, & Ho, 2019). In the programme, DHH students are co-enrolled with normal-hearing students in the mainstream classrooms. They were taught together by two teachers - one regular normal-hearing teacher and a sign bilingual Deaf or hearing teacher with proficient sign language skills. Both typically developing (TD) and DHH students are immersed in an education environment using both spoken language and sign language as the medium of instructions in class (Yiu, Tang & Ho, 2019).

Table 3. Number of TD and DHH students at different grade levels

	P1	P2	P3	P4	P5	P6	**Total**
TD students	154	184	225	105	168	127	**963**
DHH students	8	8	7	6	4	7	**40**
Total:	166	192	232	111	172	134	**1003**

Table 4. Distributions of TD subjects and their schools in different regions of HK

Region	**No. of Schools Involved**	**No. of Students**
Kowloon	**3**	**428**
New Territories West	**4**	**167**
New Territories East	**2**	**368**
Total:	**9**	**963**

In addition to the general background information, hearing-loss-related information, including DHH students' degree of hearing loss, use of hearing device, including hearing aids, cochlear implants or auditory brainstem implants were summarized in Table 5. Information about the hearing loss of the DHH students was defined in the students' audiologist's reports prepared by professional audiologists after the hearing tests. The reports were sent to schools by the Education Bureau after the DHH students received their hearing tests. In Hong Kong, the generally accepted definitions for the different degrees of hearing loss are listed in Table 6 .

The different categories of their "degree of hearing loss" were calculated based on the average hearing thresholds at the frequencies 500Hz, 1K Hz and 2K Hz of their better ear.

Table 5. Summary of the DHH students' degree of hearing loss and use of hearing devices

	Degree of Heaing Loss[b]						
Hearing Device[a]	**Unilateral**	**Mild**	**Mod**	**Mod-Sev**	**Sev**	**Prof**	**Total (%)**
ABI						5	**5 (13%)**
CI					3	13	**16 (40%)**
HA			5	3	2	4	**14 (35%)**
Unaided	2	3	/	/	/	/	**5 (13%)**
Total (%):	**2 (5%)**	**3 (8%)**	**5 (13%)**	**3 (8%)**	**5 (13%)**	**22 (55%)**	**40 (100%)**

[a] ABI=auditory brainstem implants; CI=cochlear implants; HA=hearing aids; Unaided=no hearing aids used

[b] Unilateral=hearing loss in one ear only; mild=mild hearing loss; mod=moderate hearing loss; mod-sev=moderately severe hearing loss; sev=severe hearing loss; prof=profound hearing loss (see Table 6)

Table 6. Definitions for the DHH students' degrees of hearing loss

Degree of Hearing Loss*	**Hearing Threshold (dBHL)**
Mild	25-40
Moderate	41-55
Moderately-severe	56-70
Severe	71-90
Profound	>90

* Degree of hearing loss is calculated based on the average loss at 500Hz, 1K Hz and 2K Hz of the student's better ear

As summarized in Table 5, the DHH students involved in the study had different degrees of hearing loss and used different types of hearing devices. Except that five students who had unilateral or mild hearing loss did not use any hearing device, all other students used hearing devices persistently. Four (18%) DHH students with profound hearing loss used hearing aids, 13 students (59%) used cochlear implants, and 5 students (23%) used auditory brainstem implants.

Checking with their backgrounds, 6 out of 40 DHH students were born to deaf parents while other 34 students were born to hearing parents. Considering their communication mode, all of them were able to use both sign language and spoken language to communicate with other people though their levels of competence varied. Their proficiency in sign language depends whether their parents were deaf signers or hearing parents, when did they began to learn sign language and their parents' preferred mode of communication with their children. According to Tang, Yiu, & Lam (2015), DHH students in the sign bilingual programme were able to develop the meta-linguistics awareness of both sign language and spoken language when the school environment is provided with enriched bimodal bilingual inputs.

Most of the DHH students had single disability, but three of them were confirmed to have other special needs clinically. One student with profound loss was diagnosed to have autistic features in addition to hearing loss. The other two DHH students were assessed to have Attention Deficit and Hyperactivity (AD/HD).

3.1.4 Operation and Administration

The original Chinese Grammar Assessment (CGA) profiling tool includes a total of 172 items, representing 18 grammatical categories and 48 sub-categories (see Appendix A). As the assessment targets on primary school students from P1-P6, when the items were developed, special attention was made to restrict the length of the test stimuli or answers to 5-12 characters. There are also specific features incorporated in the assessment, described as below:

3.1.4.1 Vocabulary pre-test:

All students have to complete a vocabulary test before doing the CGA. Because the assessment focuses on students' grammatical knowledge of written Chinese, it is crucial to ensure that the test results are dependent upon their morphosyntactic knowledge rather than their vocabulary knowledge (Kelly, 1996). This arrangement is to ensure that the students' performance will not be confounded by their previous vocabulary knowledge (Cannon, Hubley, Millhoff, & Mazlouman, 2016). All they need to do is to select a picture from four choices to match the meaning of the word (e.g. 烏龜 'turtle') (see Figure 2).

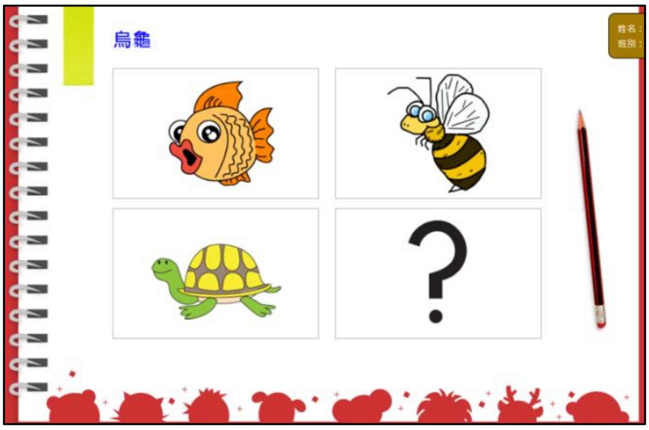


Figure 2. The format of the vocabulary test

A 75% accuracy was expected from the students. All students with a vocabulary pre-test score lower than 75% were excluded from the data analysis. The vocabularies are tested with a picture selection task. The 32 items in the vocabulary pretest, including 17 nouns, 11 verbs and 3 adjectives, are selected from the CGA test items. This group of vocabulary is repetitive in the main test and is utilized frequently in kindergarten and lower-grade primary school Chinese textbooks.

3.1.4.2 Animated video instructions:

All students are tested in front of a computer supervised by at least two investigators. Before they start the assessment, they were presented with an animated video to demonstrate how different types of test tasks should be responded to. The instructions are presented visually with no speech so that all the TD and DHH students receive the same amount of information with no barriers. After the video finished, students were given a few trial items.

3.1.4.3 Trial items:

Seven trial items are given to the students before they start with the main items to ensure that they know how the different types of questions are designed and how they should give the answers.

3.1.4.4 Randomized presentation:

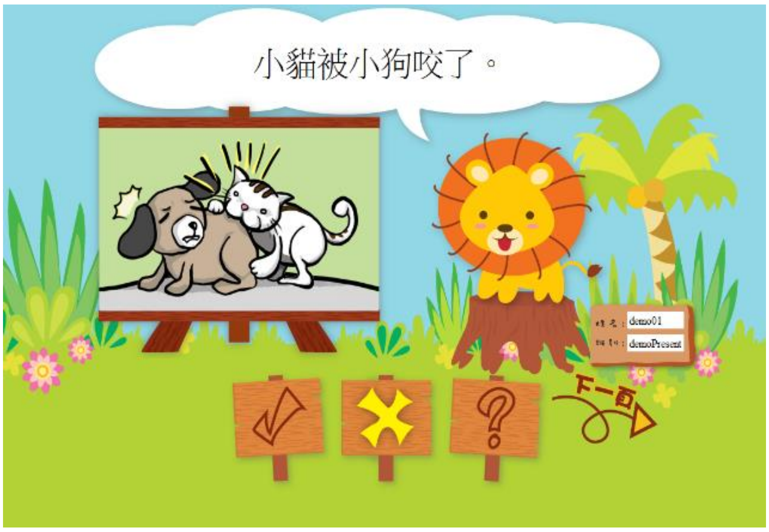
The test items are displayed on the monitor of computers or tablets. The order of presentation is randomized automatically by pre-set computer programme. Whenever a student attempts to take the test, the sequence of items will be different from the last time.

3.1.5 Task Types

CGA includes four different comprehension tasks: i) Truth Value Judgment (TVJ) task; ii) Picture Selection (PS) task; iii) Grammatical Judgment (GJ) task; and iv) Multiple Choice (MC) task. Each task covers different morphosyntactic structures and sub-categories. The distractors are designed based on the ‘form’, ‘meaning’ and ‘use’ of the specific grammatical structures and the specific linguistic properties involved. Some items are also designed with distractors following Cantonese grammar so as to see if the students are able to discriminate between the two grammatical systems, during their development from oracy to literacy.

3.1.5.1 Truth Value Judgment (TVJ) Task

The Truth Value Judgment (TVJ) task has been proven to be "one of the most illuminating methods of assessing children's linguistic competence developed in recent years" (Gordon, 1996, p.211). In the Truth Value Judgment (TVJ) task, participants were instructed to judge if the meaning of the sentence they read matched the meaning of picture on the drawing board (see Figure 3). Three choices are given. They are "Correct", "Incorrect" and "Not Sure".



小貓被小狗咬了。

姓名：demo01

[illegible]：demoPresent

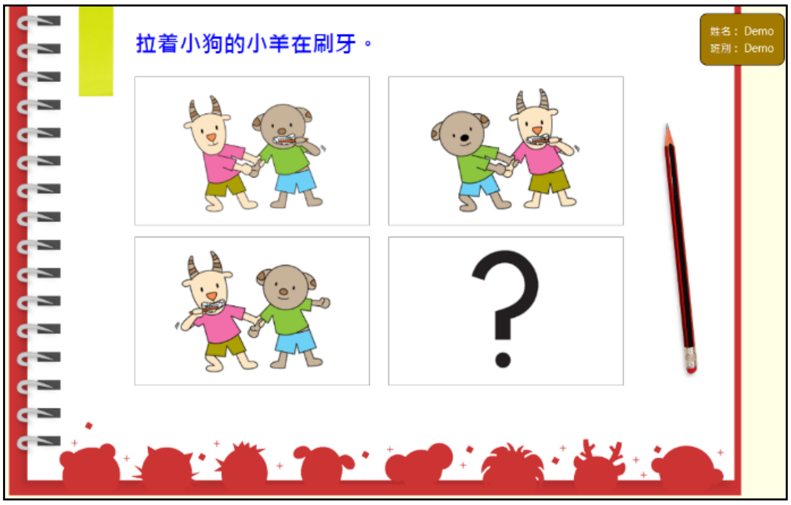
Figure 3. Interface for the Truth Value Judgment (TVJ) task

With this method used in the assessment, it is possible to evaluate if the participants understand the meaning of the sentence in a specific grammatical structure. Take Chinese passive construction as an example (see (7)), students need to understand the function and meaning of the passive marker *被 (bei6)* before they can identify the dog as the "agent" to bite the cat and the cat is the “patient” affected or bitten by the agent, i.e. the dog.

(7) 小貓 被 小狗 咬 了
cat BEI dog bite aspect
‘The cat was bitten by the dog.’

3.1.5.2 Picture Selection (PS) Task

The Picture Selection (PS) task is one of the methods most commonly used to assess children's linguistic capabilities. It is a commonly used comprehension task, especially in cases where participants failed to produce particular linguistic forms or maintain particular production contrasts (Gerken & Shady, 1996, p.125). In this task, the participants were presented with a sentence as the stimulus and asked to choose one of the three pictures that matched with the meaning of the target sentence. If they were unsure about the answer, they could choose the picture with a question mark “?” (see Figure 4). Results in this task help understand if the students are able to understand the specific linguistic structures and the functional categories involved in the items. The task can also help to see what kinds of misinterpretations the students may have (Gerken & Shady, 1996).



拉着小狗的小羊在刷牙。

姓名：Demo
班別：Demo

Figure 4. Interface for the Picture Selection (PS) task

In general, the distractors of the items are the common errors that DHH students may make. Take Chinese relative clauses as an example. Sentence (8) is a subject gap (SS) relative clause, where the embedded clause does not follow the canonical Noun-Verb-Noun strings. If children relied on the canonical word order strategy, they would misinterpret the subject gap of the embedded clause and believe that the first noun phrase, i.e., *小狗* 'dog' in (8) is the agent that *拉著小羊* 'pulls the goat'.

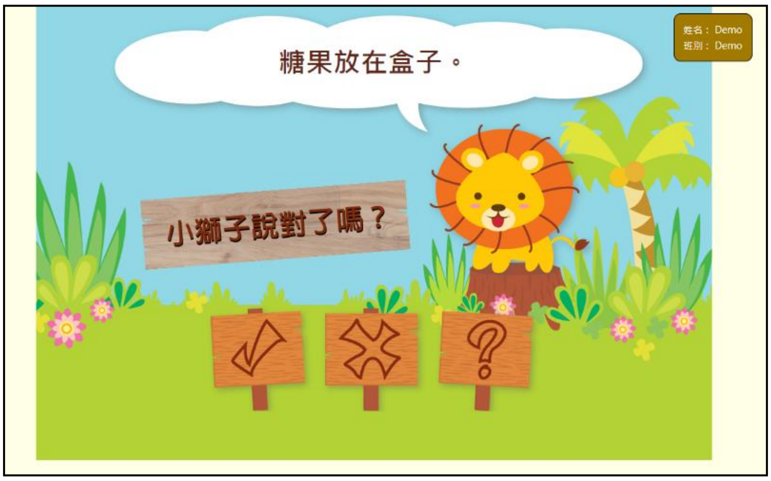
(8) 拉 著 小狗 的 小羊 在 刷牙

pull aspect dog DIK goat aspect brush teeth

'The goat that is pulling the dog is brushing its teeth.'

3.1.5.3 Grammaticality Judgment (GJ) Task

Grammaticality Judgment (GJ) is a commonly used methodology to collect linguistic data from native speakers. GJ has been used in linguistics research for nearly all different syntactic structures (McDaniel, McKee, & Cairns, 1998). GJ is frequently used in research to assess young children's grammatical knowledge (de Villiers & de Villiers, 1974). For example, the question showed in Figure 5 requires children to judge whether the sentence presented in the speech bubble is grammatically correct or not. Students can choose "Correct" or "Incorrect" based on their judgment. If the student is not certain about the answer, he or she can choose the question mark.



姓名：Demo
班別：Demo

糖果放在盒子。

小獅子說對了嗎？

Figure 5. Interface for the Grammatical Judgment (GJ) task

The sentence (9) is ungrammatical in written Chinese as the speaker does not specify the location of the sweets with reference to the box, such as *裏面* ‘inside’ or *旁邊* ‘beside’. Children with good competence in written Chinese would notice that the sentence is not grammatically acceptable.

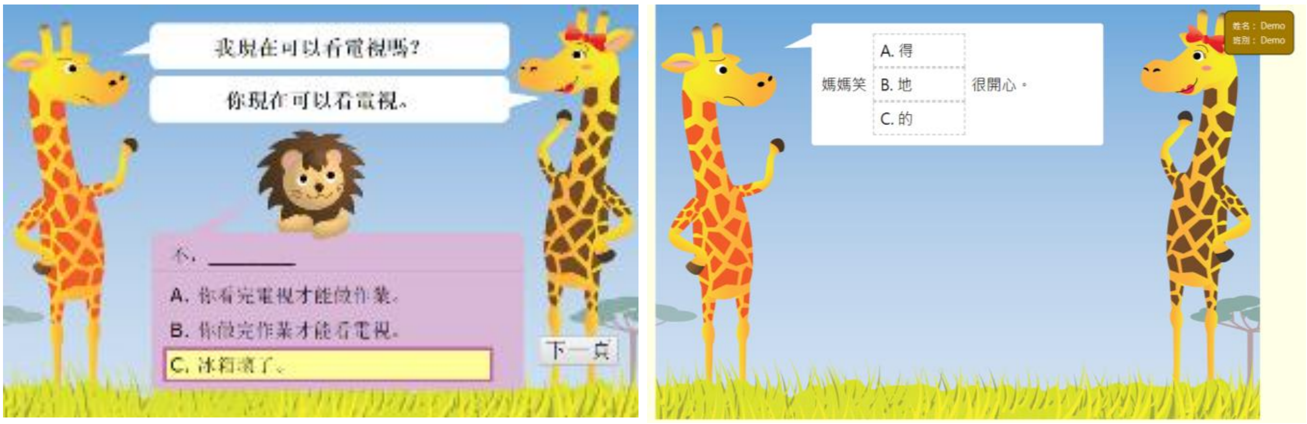
(9) * 糖果 放 在 盒子 (*ungrammatical in Chinese)

sweets put ZOI box

3.1.5.4 Multiple Choice (MC) Task

The use of multiple choice (MC) task in CGA was to assess children’s understanding of different morphosyntic knowledge in written Chinese. The multiple-choice items in CGA included different types of items such as: answering a question, filling in a blank, or completing a conversation. The choices or responses included a correct answer and 1-4 distractors, depending on the grammatical knowledge involved in the item.

CGA is used to test out if the children have a genuine understanding of some specific morphosyntactic structures or functional categories like prepositions, e.g. *從* ‘from’ and *向* ‘toward’, *wh*-words, e.g. *點樣* ‘how’ and *點解* ‘why’, and negators *不* ‘not’ and *沒有* ‘no’ (see Figure 6 for the examples).



我現在可以看電視嗎？

你現在可以看電視。

不，________

A. 你看完電視才能做作業。

B. 你做完作業才能看電視。

C. 冰箱壞了。

媽媽笑 A. 得 / B. 地 / C. 的 很開心。

Figure 6. Interface for two types of Multiple Choice (MC) tasks

3.1.6 Development of Two Alternate Forms of CGA

As mentioned above, the initial version of CGA consists of 172 test items in 18 major categories and 48 subcategories for profiling students’ grammatical knowledge. Most sub-categories included 4 items, and some included 2 items. This initial version aims to collect a relatively comprehensive linguistic profile of individual students, allowing a more in-depth analysis of DHH students’ performance. In order to achieve this objective, quite a large number of items were included to cover a wide range of grammatical categories in written Chinese, resulting in a development of an overly long assessment for daily educational and clinical applications. Especially for junior primary students, it was difficult for them to keep their attention for the full assessment except that it had to be separated into a few sessions. In addition, as we only have one single version of this profiling tool, no retest can be done within a short period of time because of the possible learning effects. Therefore, it would be hard for the teachers and clinicians like speech therapists to use the assessment to review students' progress after their interventions. Developing two CGA short tests can eventually be used more effectively in various educational and clinical applications.

Before the two alternate forms of CGA can be established for the two short tests, we must ensure that the current test items are valid and reliable for both typically TD and DHH students. To have a thorough review of the items, content validation and psychometric review are of crucial importance. The former procedure is to ensure the items are representative and relevant to the targeted latent traits of students, and the design of the items are appropriate for the testing of TD and DHH students' grammatical knowledge in Chinese. The latter is to collect objective evidence to confirm that the items selected for the alternate forms are psychometrically valid and reliable.

3.2 Phase One: Content Validation for CGA

Content validation is one of the important steps to establish a valid assessment. "Though the usefulness and reliability of using expert judgments as a means of analyzing the content or difficulty of test items in language assessment has been questioned for more than two decades. Still, groups of expert judges are often called upon as they are perceived to be the only or at least a very convenient way of establishing key features of items" (Alderson & Kremmel, 2013, p.535). As mentioned, all the items were reviewed by three experts in Chinese linguistics before they were finalized and used in field testing. To validate the assessment for a broader scope of usage, especially in support of the work of educators and speech and language professionals, it is essential to have further review conducted by related experts, who will practically use the two short tests to support DHH students' development. To match with the objective, in this round of expert review, instead of linguistics experts, professional teachers and speech therapists who possessed subject knowledge in language testing and Chinese Language education for both TD and DHH students were invited as the Subject Matter Experts (SMEs) to conduct the content validation review for the items and the assessment as a whole.

With the provision of the dataset provided from the Centre for Sign Linguistics and Deaf Studies, The Chinese University of Hong Kong, in order to develop two alternate forms, and then the two CGA short tests for primary school students, a series of psychometric review and validation procedures were conducted. In the following paragraphs, we will discuss the methodology used in the review with reference to other similar studies.

3.2.1 Procedures for Expert Review

"Content validity, a critical step in the test development and validation process, refers to the degree to which elements of an assessment tool are representative of the construct of interest and appropriate for a given population" (Hubley & Palepu, 2007, p.47). The evidence of test validity includes not only the review and endorsement of test items, but also the test title, instructions, display and response formats and scoring methods, etc. (Cannon & Hubley, 2014; Hubley & Palepu, 2007). As remarked in the literatures, there is no set number of SMEs required but a range of 5-10 SMEs was generally recommended (Lynn, 1986). "The more the experts, the greater the confidence in the ratings and the easier is to detect rater outliers" (Hubley & Palepu, 2007, p.47).

In this study, in order to examine the content validity of the 172 items of the CGA profiling tool, a panel of 10 subject matter experts (SMEs) were formed. Among the panel members, there were five professional speech therapists with an average of 15.3 years (a range of 6-25 years) of experience in the field, and an average of 7.5 years of working experience for DHH students. The other five panel members were teachers teaching Chinese Language for both TD students and DHH students in primary schools. They had an average of 11.8 years (a range of 5-12 years) of teaching experience with an average of 8.4 years of teaching for DHH children. All of them had no direct involvement in the development of CGA.

A CGA review platform was set up for the SMEs to conduct the review and give their ratings and comments online. To help them understand the objectives of the review of CGA, including the reasons for the development, the major objectives, the design of the assessment platform and the potential applications of CGA, etc., the platform started with some explanatory notes, explaining how different areas of contents should be reviewed. The SMEs were then guided to fill out the content validation questionnaire for our further analysis (see Appendix C). In the review, the results in terms of SMEs' ratings and Content Validity Index (CVI) were used for separate reasons:

i. The ratings were used to tap the degrees of endorsement for individual items or areas of CGA development, and
ii. The CVI was to check the degrees of consensus among the ten subject matters experts.

3.2.2 Rating for Individual Items or Areas of Development

With reference to Cannon & Hubley (2014), the SEMs were asked to review the contents of CGA. Different 5-point Likert scales were used to rate the representativeness of the selected grammatical categories and the 172 items' relevance, and appropriateness. As suggested by Østerås et al., (2008), the 5-point Likert scales are more reliable than the 4-point scales. Therefore, the representativeness of the 18 grammatical categories was rated on a 5-point

Likert scale with the scoring system as follows: “1” = very poor representativeness, “2” = poor representativeness, “3” = fair representativeness, “4” = high representativeness, and “5” = very high representativeness. The 5-point scales for the items’ relevance and appropriateness were developed in a similar way: “1” = very poor relevance/appropriateness, “2” = poor relevance/appropriateness, “3” = fair relevance/appropriateness, “4” = high relevance/ appropriateness, and “5” = very high relevance/appropriateness. As the norms for the two CGA short tests were developed based on the performance of typically developing (TD) students, the items were reviewed according to the perspective and needs of the general population rather than that of the DHH students.

Table 7. Questions about the appropriateness of the administration and operations of CGA

CGA operational elements		The appropriateness of the design*	Recommendations in this regard, if any
1.	Operating as a web-based online assessment	□1 □2 □3 □4 □5	
2.	Displaying items randomly by the computer – every time in a different order	□1 □2 □3 □4 □5	
3.	Students can change their answers before submission	□1 □2 □3 □4 □5	
4.	Using an animated video to explain how to answer different types of questions	□1 □2 □3 □4 □5	
5.	The contents and the illustration of the video	□1 □2 □3 □4 □5	
6.	Doing trial items before doing the test items	□1 □2 □3 □4 □5	
7.	Receiving a vocabulary test before doing CGA	□1 □2 □3 □4 □5	
8.	The number of vocabularies in the vocabulary test	□1 □2 □3 □4 □5	

* A 5-point scale is used ‘1’ = very inappropriate, ‘2’ = inappropriate, ‘3’ = fairly appropriate, ‘4’ = appropriate, and ‘5’ = very appropriate

Besides reviewing individual grammatical categories and the 172 test items, the overall design of the assessment and its appropriateness to TD and DHH students were also evaluated by the panel with guided questions as listed in Table 7. For all sections of the review, some space for open remarks was provided in the different checklists so that the SMEs could provide more detailed explanations on their ratings and provided some further suggestions for the items, their design and the operation elements of the assessment (see Table 8).

Table 8. Question about the overall design of CGA

Please answer the questions below:

1. How appropriate is the title “Chinese Grammatical Assessment (中文語法評估)”?
□ very inappropriate □ inappropriate □ fairly appropriate □ appropriate □ very appropriate

2. How appropriate is the mode of operation of the assessment?
□ very inappropriate □ inappropriate □ fairly appropriate □ appropriate □ very appropriate

3. Are the selected 18 grammatical structures of CGA representative of the literacy development of primary school children?
□ very poor representativeness □ poor representativeness □ fair representativeness
□ good representativeness □ very good representativeness

4. Are test items of CGA suitable for testing Chinese grammatical knowledge of typically developing children?
□ not suitable □ not really suitable □ fairly suitable □ suitable □ very suitable

5. Are test items of CGA suitable for testing Chinese grammatical knowledge of deaf or hard-of-hearing children?
□ not suitable □ not really suitable □ fairly suitable □ suitable □ very suitable

A commonly used methodology, the Content Validity Index (CVI) (Lynn, 1986) was adopted in the study to evaluate SMEs’ agreement on the ratings of various operational elements of CGA. After the panel’s review, CVIs for all items and questions were calculated to quantify SMEs’ level of endorsement as a group. Two levels of statistics were calculated, ratings from an individual item level and overall ratings for the assessment (Hubley & Palepu, 2007). To ensure a more concrete endorsement, a rating of “4” or “5” was considered a positive endorsement by an SME whereas a rating of “1”, “2” and “3” were basically treated as a non-endorsement in the present study. Therefore, for every rating from an SME >3, the value of CVI=0.1; for two SME’s ratings >3, CVI=0.2, etc. As we had 10 SMEs in the expert panel, the maximum CVI value is 1.0.

3.2.3 Content Validity Index

CVI values for the assessment as a whole are defined as “the average proportion of items endorsed by the SMEs” (Hubley & Palepu, 2007, p.49), and so it is calculated by averaging all items’ CVI values. As suggested in Lynn (1986), a minimum of 8 out of the 10 SMEs’ endorsement (i.e. CVI > 0.80) was required to achieve a significant evidence (α=0.05) to justify the content validity of the items or the elements of CGA. Those items or elements not endorsed by 8 out of 10 SMEs were examined further to determine if appropriate revisions were required.

3.3 Phase Two: Psychometric Review of the Items

More and more language assessments are developed based on Item Response Theory. Aryadoust (2022) in the review of different studies, has identified different major areas of research and focal investigations, covering a wide range of language constructs such as reading, speaking, listening, vocabulary and grammatical knowledge (see also Min & Aryadoust, 2021). Rasch model (Rasch, 1960) was used in this study to analyze the dichotomous data collected from the students based on CGA, in which the same scoring system (i.e. “0” for incorrect answers and “1” for all correct answers) was used for all the items in the assessment. The advantage of Rasch analysis is that the person and item statistics can be assessed together and highlighted on the same scale, which provides useful information to observe the construct validity of the measurement (Efeotor, 2014). In addition, the analysis based on item response enhances the reliability of the results. The results provided concrete evidence for item responses and helped select good-fit items for the two CGA short tests.

Differential Item Functioning (DIF) analysed in Rasch helped identify items that are biased or disadvantaged to a particular population in CGA. In this study, though the norms were developed based on typically developing students, the test was meant to develop also for students with hearing loss. To facilitate an objective review of the psychometric properties of the existing items used in CGA, Rasch analyses were conducted for two conditions: TD data only and "TD plus DHH" data. No DHH data would be analyzed alone due to small sample size. The analyses included the fit statistics for person ability and item difficulty, separation reliability of person and item, Wright maps, dimensionality and analysis of Differential Item Function (DIF).

3.3.1 Reliability and Internal Consistency

For every Rasch analysis, reliability measures concerning person ability and item difficulty are provided for the psychometric evaluation. High values in item and person separation reliability indicate that the test is reliable to discriminate between the persons participating in the test based on their ability, and it is also effective in discriminate items with different difficulty (Efeotor, 2014, p.213). In this study, both person and item reliability > 0.8 and separation >2 were expected. Besides, Cronbach's alpha was also conducted for the review of CGA's internal consistency.

3.3.2 Fit Statistics

Having analyzed properties of individual items distinctively, the Rasch model is very effective in determining if the items fit well with the model. The analysis is especially useful in this study, which aims to develop two alternate forms with good-fit items for the development of two normative short tests (Flanagan, 1951). The "infit" and "outfit" statistics are significant information to see if the observed response corresponds well with those predicted by the model.

In the analyses, for those items that were found not well-fitted into the model, they were scruitnated or deleted unless there were other valid evidence to support the retention.

Outfit and infit statistics in terms of mean-squares (MNSQ) are the major indicators to help determine which persons and items data should be kept for further analysis (Linacre, 2002; Bond & Fox, 2007). For different types of measurements, there are specific recommendations for the range of item INFIT MNSQ and OUTFIT MNSQ (Wright & Linacre, 1994). As the average of calculated mean-squares is 1.0, a range between 0.5 to 1.5 is considered an acceptable range of INFIT/OUTFIT MNSQ values for a productive construction for measurements (Wright & Linacre, 1994; Bond & Fox, 2007; Boone, Staver, & Yale, 2014). Following the aforementioned recommendations, the requirement of 0.5 < INFIT/OUTFIT MNSQ < 1.5 was adopted for the current study as a reference to review the data fitness of person ability and item difficulty.

The range of Z-Standardized values (ZSTD) is expected to be -2.0 to 2.0 (Bond & Fox, 2007). As suggested by Linacre (2019), when MNSQ values are within the acceptable range between 0.50-1.50, no specific checking is required for the ZSTD values (Dragounova, 2018). The requirement for Point Measure Correlation (PTMEA-CORR) of individual items is basically a positive value, showing all items are correlated to the assessment as a whole. In this study, we looked into the INFIT/OUTFIT MNSQ, Z-standardized values (ZSTD) and Point Measure Correlation (PTMEA-CORR) to check for the fitness of the items (Boone, Staver, & Yale, 2014). Any items that failed to fulfill all the three criteria listed in

Table 9 would be deleted for a more reliable review of the assessment (Abul Aziz et al., 2014).

Table 9. Three criteria for checking item fitness for a test (Boone, Staver, & Yale, 2014)

Statistics	Aim	Fit Indices	Interpretation
Outfit mean square values	Fitness of items	0.5-1.5	Items should be changed or removed when all three criteria are out of the fit range
Outfit z-standardized values (ZSTD)		-2.00-2.00	
Point Measure Correlation (PTMEA-CORR)	Item polarity	0.4-0.85	

Before reviewing the items' fitness, fit statistics were conducted for the persons first. Misfitted person data were removed to avoid problems of "underfit meaning there is too much unexplained variance (or noise) in the data, and…overfit meaning the model overpredicts the data causing inflated reliability statistics" (Boone, Staver, & Yale, 2014, p.166). The adopted procedure was to ensure that the items selected for the two alternate forms of CGA were valid and reliable. They were good-fit items for the assessment of the targeted latent trait, that is, grammatical knowledge in written Chinese.

3.3.3 Differential Item Functioning (DIF)

The development of CGA is based on validation data from TD students. With reference to their ability in Chinese grammatical knowledge and the projected norms, it aims to provide an assessment platform to assess DHH students' performance and understand their needs for literacy development. To ensure testing fairness to both DHH as well as TD students as far as possible, the design of the test items should have no bias toward either populations (Efeotor, 2014). Differential Item Functioning (or DIF) was conducted with TD and DHH data together, which help identify test items that may be potentially bias or disadvantage to either DHH or

TD students. For example, test items including the concept of sound may be unfavourable to DHH students. The results from the Mental-Haenszel test were used for the DIF analysis for the dichotomies data collected from CGA. According to the recommended guidelines of Zwick, Thayer & Lewis (1999), items with absolute values of DIF Contrast >2 and a probability *p*≤ .05 from the Mental-Haenszel test results would be flagged for scrutiny. Whether the items are (dis)advantaged to TD or DHH students would also be investigated and considered an exclusion from the two alternate forms. According to Scott and the team in 2009, dichotomous dataset should better be >1000 data for each sub-group for a robust DIF analysis, so the conduction of the DIF test here is considered only a trial (Linacre, 2012).

3.3.4 Dimensionality

The Chinese Grammatical Assessment (CGA) was developed to assess students' grammatical knowledge in written Chinese as the targeted latent trait for the measurement. In principle, all items included in CGA should fall into the same dimension. Dimensionality is inevitably an important element to consider when a language assessment is to be developed. Whether a unidimensional or multidimensional Rasch model should be used to analyze a language test is always controversial regarding each model's practical benefits and limitations (Reise, Cook & Moore, 2014). According to Min and Aryadoust (2021), multidimensionality and unidimensionality "were almost equally adopted across research on listening, reading, speaking and writing, whereas an overall dominance of the unidimensional framework was found in vocabulary and grammar assessment" (p.7). Grammar tests are mostly analyzed from a unidimensional perspective, assuming that grammatical knowledge falls under the same construct or latent trait (Efeotor, 2014). A similar assumption was adopted for the analyses of the two CGA short tests.

Sumintono and Widhiarso (2015) provided the criteria of unidimensionality based on the “raw variance explained by measures” from the standardized residual variance. The value of “raw variance explained by measures” which is higher than 20% is acceptable, higher than 40% is good, while higher than 60% is excellent. Meanwhile, the ideal value for the “unexplained variance” should not exceed 15% (see Table 10). When the items of the two alternate forms were initially confirmed, the dimensionalities of the two short versions of CGA were then reviewed.

Table 10. The standard for dimensionality measures in Rasch Analysis

Statistics	**Aim**	**Value of raw variance**	**Interpretation**
Dimensionality	Check if the model of the measurement should be unidimensional	***Explained variance***	
		>20%	Acceptable
		>40%	Good
		>60%	Excellent
		Unexplained variance	
		≤ 15%	No other dimension

Source: Sumintono and Widhiarso (2015) quoted by Saidi and Siew (2019, p.544)

3.4 Phase Three: Item Selection and Validation of the Two Short Tests

3.4.1 Selection Criteria

After different psychometric reviews of the 172 items of CGA, some items would be selected for the two alternate forms considering the following criteria: i) INFIT/OUTFIT MNSQ should be within the acceptable range of 0.50-1.50; ii) the ratings from SMEs >4.0 (out of 5.0) and the projected Content Validity Index (CVIs) > 0.8 (out of 1.0) regarding the item’s relevance and appropriateness for assessing students’ Chinese grammatical knowledge; iii) the items were not biased to either TD and DHH subgroups in the results of Differential Item Functioning

(DIF); and v) an equivalent level of difficulty was projected from the two alternate forms.

Upon completion of the analyses mentioned above, the final step of item selection for the two alternate forms would be proceeded. For those items that did not fit well with the set criteria were identified and flagged for further investigation accordings to the different results of psychometric reviews. Scrutinization of individual items concerning the original design was conducted to determine which items should be selected for the two forms or excluded from the final lists of items. Once the two alternate forms were confirmed, different reliability and validity measures were then conducted to further validate their psychometric properties of the two projected CGA short tests before establishing the norms for them. The reliability and validity measures used in the study will be discussed in the following sections.

3.4.2 Reliability Measures

Reliability is concerned with the extent to which an assessment is consistent in repeated measurements. Good reliability is a foundation for achieving test validity. In this study, a series of reliability measures were used to review the two short tests of CGA including internal consistency, person and item separation reliability, alternate forms reliability and test-retest reliability according to the TD and DHH data extracted from the database established by the Centre for Sign Linguistics and Deaf Studies, The Chinese University of Hong Kong from 2015-2019 (Tang et al., 2022), and a newly collected dataset from a school adopted the Sign Bilingualism and Co-enrollment in Deaf education Programme in 2022 (see results reported in Chapter 7 and Chapter 8).

3.4.2.1 Person/Item Reliability and Internal Consistency

As mentioned in Section 3.3.1, the criteria set for the item/person reliability in Rasch analysis was >0.8 and the item/person separation index was be >2.0 for this study. Internal consistency reflects the uniformity of test items, but it is also a prerequisite of construct validity. In this study, Cronbach's alpha was also conducted to check the two short tests' internal consistency. The values of Cronbach's alpha ranges from 0 to 1.0. A good reliability value of >0.7 (Saidi & Siew, 2019) was expected for the two alternate forms of CGA.

3.4.2.2 Alternate Forms Reliability

The alternate forms reliability is to review the results of the two measurements from the same group of raters or test participants (Holmefur et al., 2009). In this regard, the test results by the two alternate forms should be highly correlated for the same group of subjects. Besides, the results of a heterogeneous sample group assessed by the two short tests should have no significant difference between each other. In view that the correlation coefficient itself is not able to pick up the discrepancy in variances of the students' results assessed by the two alternate forms of CGA, the Intra-class Correlation Coefficient (ICC) was used for the analyses, taking into account both the association between the two sets of test results and their variance of the data.

There are different modes of analysis for ICC. A different "definition" we selected would bring to a significantly different result from the analysis (Koo & Li, 2016). As the raters or participants involved in this study were the only subjects of interest in the analyses, no randomized samples were involved, and thus the two-way mixed-effects model was adopted (McGraw & Wong, 1996).

There are two types of analyses for ICC, “[a]bsolute agreement concerns if different raters assign the same score to the same subject. Conversely, consistency definition concerns if raters’ scores to the same group of subjects are correlated in an additive manner” (Koo & Li, 2016, p.158). In this study, the definition of “absolute agreement” was selected as the main concern of the alternate forms reliability for CGA. The extent to which the scores of the two forms, CGA-A and CGA-B, were equivalent to each other was the main concern of the analyses.

As we only took single measurements for all individual subjects in this study, the type of analysis was thus simply defined as “single measures”. Following the discussion mentioned above, a single-measure, absolute-agreement, two-way mixed-effects ICC model would be used for the alternate forms reliability between the two CGA short tests. In this study, the SPSS version 2.7 was used for the ICC analyses between data from CGA- and CGA-B.

3.4.3 Validity Measures

Validity is one of the most important qualities of a test. There are essentially three approaches to test validation including: 1) content validation, concerning the relevance of test contents to the characteristics being measured, 2) criterion validation, concerning the hypothesized relationship of the test with external criteria, and 3) construct validation, concerning the internal structure of the test (Hammond, 1995).

3.4.3.1 Content Validity

As discussed earlier in this chapter, an expert panel with ten Subject Matters Experts (SMEs) was formed to help review the representativeness, relevance, and appropriateness of the assessment and the 172 items of CGA profiling tool. Having received their ratings, togther with the projected Content Validity Index (CVI), the resultant ratings and CVIs of the two alternate forms were further investigated according to the selected items of the two alternate forms of CGA. This serves as a collection of significant evidence from subject experts regarding the content validity of the two finalized lists of items and eventually the two CGA short tests.

3.4.3.2 Criterion Validity/Convergent Validity

Criterion validity can be in a form of predictive validation or concurrent validation (Hammond, 1995). Grammatical knowledge or morphosyntactic understanding is considered a good predictor of DHH students' reading skills (Kelly, 1996). It is crucial to see if CGA can also predict students' reading ability or academic performance in Chinese. However, there is no standardized assessment available for the Chinese grammatical knowledge of students. No gold standard is established for the assessment of related construct.

As an alternative, the research team collected data from school examination in Chinese Language for both TD and DHH students in a school, specifically the reading and writing performance in the examination. This serves as a test for convergent validity of the two alternate forms of CGA. For the validity check for data of DHH students, the results of DHH students in a normative academic assessment, namely the Learning Achievement Measuring Kit (LAMK; Education Bureau, 2008, 2014) was collected. LAMK is a well-known assessment in academic performance of students with special needs, for identifying their learning needs and progress in different major subjects. The test includes three subjects, Chinese Language, English Language and Mathematics, and the test results of LAMK can be considered as a gold standard for academic performance in Chinese Language. The relationships between CGA scores and DHH students' results in LAMK can be considered as a proof for Criterion Validity though it can only represent the results in Cantonese-speaking DHH subjects.

3.4.3.3 Construct Validity

Collective evidence is required to confirm the construct validity of the measurement (Efeotor, 2014). Examining the internal consistency and reliability of the two CGA equivalent lists are the necessary condition to support the two short tests' construct validity (Hammond, 1995). As discussed in Section 3.4.2.1, reliability scores based on person and item reliability were expected to be >0.8 with the values of separation index >2.0 based on Rasch analysis. Cronbach's alpha was expected to be >0.7. With reference to Efeotor (2014), checking with the results from the Wright maps and item outfit MNSQ also helped to see if the items of the two lists were well-fitted with the model and testing for the same construct.

Another measure conducted for the review of construct validity of CGA was the assessment of its known-groups validity. Known-groups validity focused on the two or more groups, which were known to have or logically should have different levels of in the targeted latent trait (Davidson, 2014). In view that students at primary schools should have continuous development in their Chinese grammatical knowledge during their six-year learning process in Chinese Language in schools. Therefore, a significant main effects of grade levels on students' CGA test scores was expected. The analyses would be conducted by One-way ANOVA and post-hoc test for pairwise comparisons between students' CGA scores at different grade levels. The investigation helped to provide evidence for the construct validity of CGA.

3.4.4 An Interim Discussion

The process of item selection is to create two balanced sets of good-fit items for the two alternate forms of CGA so that the two respective short tests can be used for the assessment of students' Chinese grammatical knowledge interchangeably. For each alternate form of CGA, items were selected from the 172 items, developed from 18 categories and 48 sub-categories of grammatical or morpho-syntactic knowledge in written Chinese. It was intended to select one-item from one grammatical sub-category in order to cover a wider scope of linguistics properties in the two short tests. Before the two alternate forms were finalized, content validation by the expert panel was required to check for representativeness of the grammatical categories and included in CGA, and the relevance and appropriateness of the items. Then a series of reliability and validity measures were conducted for the validation of the two equivalent lists of items.

After the above procedures were completed, the two short tests, CGA-A and CGA-B were formed. Under a genuine testing condition, they were used to assess a new group of subjects with both TD and DHH students for further validation. One short test lasted for around 15-20 minutes, which was relatively easy for primary school students to manage, compared to the original 172-item CGA profiling tool. The data from the newly tested subjects could provide additional information to re-assure that the two tests were well validated with good reliability and validity. Some data were also collected for additional tests on reliability and validity, including the repeated testing data within 1-3 weeks for the the review of test-retest reliability, and the academic data of students' Chinese Language examination or normative assessment for establishing the two tests' convergent validity.

3.4.5 Reliability and Validity Measures for a New Dataset

As mentioned above, to further assess the reliability and validity of the two newly established CGA short tests, a new set of 102 TD and 27 DHH data was collected from a primary school adopted the Sign Bilingualism and Co-enrolment in Deaf Education (SLCO) Programme.

The SLCO Programme was first established at a mainstream kindergarten in 2006, and gradually extended to primary and secondary education in Hong Kong, using a whole school approach to inclusive deaf education (Tang et al., 2023). The primary school has started implementing the SLCO Programme in school since 2016. Since then, a critical mass of 3-6 DHH students were admitted to the school each year, and they were all integrated in the regular classes in groups, rather than distributed into different classes individually like general mainstream schools. All DHH students at the same grade level were grouped in one class, and learned together with their typically developing classmates for all lessons, following the same curriculum for all subjects, including Chinese Language. In the 6 SLCO classes at 6 different grade levels, a small group of deaf students were integrated into regular classrooms with their hearing peers, co-taught by a regular subject teacher and a sign bilingual teacher, Deaf or hearing, using both sign language and spoken language (in either oral or written form) to conduct the lessons (Yiu, Tang, & Ho, 2019). Besides some lessons for students' learning of Putonghua, Hong Kong Sign Language, Cantonese and written Chinese were the major instructional media used for the Chinese Language lessons in the SLCO classes.

3.4.5.1 Participants

A group of the TD and DHH students studying from P1-P6 were tested by both CGA-A and CGA-B to support further validation of the two short tests, and better understand the DHH students' grammatical development in written Chinese. Altogether, 112 TD students from P1 to P6 were involved in the study. After checking the results from the vocabulary pre-test, 10 TD subjects with the test scores below 75% were excluded from the study, leaving 102 TD subjects for further analysis. Table 11 summarizes the number of TD and DHH data from the new dataset for further validation of CGA short tests.

Table 11. Number of TD and DHH data from the new dataset for further validation of the two CGA short tests

	P1	P2	P3	P4	P5	P6	**Total**
TD students	12	15	17	22	19	17	**102**
DHH students	5	4	6	3	4	5	**27**
Total:	17	19	23	25	23	22	**129**

Table 12. Summary of students' degree of hearing loss and their use of hearing devices

	Unilateral	**Mild**	**Mod**	**Mod-Sev**	**Sev**	**Prof**	**Total**
ABI						3	**3**
CI					2	12	**14**
HA			2	3	1	3	**9**
Unaided	1	/	/	/	/	/	**1**
Total	**1**	**0**	**2**	**3**	**3**	**18**	**27**

[a] ABI=auditory brainstem implants; CI=cochlear implants; HA=hearing aids; Unaided=no hearing aids used
[b] Unilateral=hearing loss in one ear only; mild=mild hearing loss; mod=moderate hearing loss; mod-sev=moderately severe hearing loss; sev=severe hearing loss; prof=profound hearing loss

For DHH students, hearing-loss-related information was also collected. Table 12 summarizes the students' degrees of hearing loss and their use of different hearing devices. Most of the DHH students in the study used cochlear implants (*N*=14; 51.85%) and hearing aids (*N*=9;

33.33%). A few of them received operations for the auditory brainstem implant (*N*=3; 11.11%). Among the 27 DHH students, 21 (77.78%) of them had a severe or profound hearing loss. Two of them (7.41%) had a moderate hearing loss. One student had a unilateral hearing loss (having normal hearing ability in one ear and a significant loss in another ear) without using any hearing device. As a summary, most DHH subjects in the part of study suffered from a severe-to-profound level of deafness.

3.4.5.2 Assessing Validity and Reliability of the Two CGA Short Tests

To further validate the two CGA short tests, a series of reliability and validity measures were conducted using the new dataset. Rasch analyses were administered for the two sub-sets of data collected from CGA-A and CGA-B separately. Results regarding the two tests' separation reliability, internal consistency, fit statistics will be reported in Chapter 8. Data from TD and DHH subjects were also analyzed separately in some measures though the sample size of DHH subjects were small.

3.4.5.3 Reliability Measures

Reliability measures, including alternate forms reliability and the test-retest reliability were conducted based on data from the DHH and TD students in the 6 SLCO classes. Therefore, all students were tested with the two CGA short tests in 2 sessions. Statistical analyses were conducted to review if the students performed similarly in the two short tests for their alternate forms reliability, and in the repeated testing situation for their test-retest reliability. As expected, the test scores collected from two above-mentioned conditions should be highly correlated and comparable with each other (Wilson, 2001). To have a thorough review, the Intra-class Correlation Coefficient (ICC) was used to evaluate both the alternate forms reliability and test-retest reliability according to the results of the CGA short tests.

Test-retest reliability is evaluated by testing subjects on repeated occasions. As mentioned that the TD and DHH students in this study were tested twice for both CGA-A and CGA-B. The re-test was administered within 1-3 weeks after their first test. The aim is to check for stability of test results over time as one of the parameters to evaluate the reliability of the assessments (Holmefur et al., 2014). The Intra-class Correlation Coefficients (ICCs) were used based on a single-measure, absolute-agreement, two-way mixed-effects model. The analyses were conducted for different subject groups (TD only, DHH only and a combined group of TD and DHH subjects) based on their raw scores.

3.4.5.4 Validity Measures

To assess the convergent validity of the two CGA short tests, one method is to examine correlations between CGA performance and some existing similar measures for the targeted latent trait such as grammatical knowledge in Cantonese, reading and writing abilities, academic performance in Chinese Language, etc. The evaluation is to collect convergent evidence that supports the valid interpretation of assessment scores obtained from the two CGA short tests (Gioia, Espy, & Isquith, 2003).

Grammatical knowledge is a significant factor affecting reading comprehension of both TD or DHH students. However, there is no standardized measure on reading comprehension. Therefore, in this study, students' academic performance in Chinese Language (including reading comprehension and writing skills in written Chinese) would be used as another measure to explore the convergent validity of CGA. As suggested by Stinson and Antia (1999), both normative academic status and classroom academic status of the students were considered, depending on the available data for the two subject groups. Classroom academic status refers to students' performance compared to their classmates, while normative academic status refers to students' performance with based on standardized academic assessments. In this study, students' academic scores got from their year-end or final school examination in Chinese Language represented their classroom academic status.

In view that the examination papers were different for different grade levels, no fair comparison can be made between students from different grade levels according to their raw scores. To facilitate further statistical analysis, the percentile ranks were calculated. For each grade levels, students' raw scores were converted to percentile ranks. In this case, the performance of individual students was represented by their percentile rank with reference to the results of his or her peers at the same grade level. Students with a better examination score was positioned at a higher percentile rank at their grade level.

For DHH students, besides school examination results, they were also tested by a standardised academic assessment, the Learning Achievement Measuring Kit (LAMK; Education Bureau, 2008, 2014) for their normative academic attainment in Chinese Language. As a general policy defined by the Education Bureau, all DHH students under the support of the schools have to receive the test to help report the students' progress and their educational needs to the government. LAMK was piloted in 2006, and then revised as LAMK 2.0 and standardized in 2008 with the Cronbach's alpha, α=.88 (Education Bureau, 2008). LAMK was then further upgraded to LAMK 3.0 with a Cronbach's alpha, α=.90 (Education Bureau, 2014).

The standard scores of LAMK serves as a gold standard for academic assessment. A good correlation between students' CGA scores and the standard scores of LAMK suggests a positive convergent validity of CGA. In this study, the relationships between DHH students' Chinese Language assessment by LAMK and CGA scores in both short tests were assessed. The results provided valuable evidence that supports the convergent validity of CGA.

Once the reliability and validity of the two short tests were confirmed, their individual norms would be established according to the data collected assessment from the typically developing students.

3.5 Phase Four: Developing the Norms for CGA

The development of the two CGA short tests, namely CGA-A and CGA-B, aims to review how DHH students perform with reference to the normative standards of typically developing (TD) students. There are multiple advantages of this development in different educational and clinical applications. Firstly, with two shorter versions of CGA, students can complete the assessment with less time and better attention. Secondly, the assessment with established norms can be used more effectively in daily education or clinical practices in identifying the needs of DHH students. In addition, the availability of two equivalent short tests helps to track students' development in Chinese grammatical knowledge interchangeably at different time points. Lastly, a shorter version allows more rooms for further inclusion of representative new items in the two tests (Ng, 2014).

The norms of the two short tests for Chinese grammatical knowledge would be developed based on the raw scores of typically developing (TD) students studying at different grade levels.

Before establishing the norms, a prior check for normality of data distribution conducted by the Shapiro Wilk test, which is commonly used for checking normality of different datasets. A significant *p*-value <.05 of the test result represents that the data do not fulfil the normality assumption. The norms for the two short tests would be calculated by SPSS version 27 expressed in terms of percentile ranks since the normality assumption was not supported by the Shapiro Wilk test (see Chapter 9).

After setting up the norms, a crucial question we need to consider answer is "below which percentile rank that a student should be considered as having a delayed development in CGA". With reference to a renowned language assessment, the Clinical Evaluation of Language Fundamentals-Fifth Edition (CELF-5; Wiig, Semel, & Secord, 2013), a percentile rank of 16 or below, which is equivalent to a -1 standard deviation in a normal distribution is classified as a "below average" performance. According to this classification, a percentile rank between 17 and 83 was considered "average" performance and a percentile rank ≥84 was considered "above average" performance. To investigate if this classification is effective in identifying and understanding the needs of DHH students, it would be applied to the newly collected data from the Sign Bilingualism and Co-enrollment in Deaf Education (SLCO) Programme for an exploratory analysis. In the following section, we will explain how the case study was conducted.

3.6 Phase Five: A Case Study with a Group of DHH Students

The last phase of the study is a case study based on a new set of data from a group of TD and DHH students as mentioned in Section 3.4.5.1. Following the classification as defined in Section 3.5, DHH students were grouped into three groups: "above average performance", "average performance" and "below average performance" based on the CGA performance of the 27 DHH students in the two tests (see Table 11 and Table 12 for information of the subjects included in the case study). Then some further observations and investigations were conducted to explore how CGA score related to their background, their deafness-related factors such as degree of hearing loss and their academic performance. More importantly, the case study is to see if the norms are helpful in identifying DHH students with a relatively delayed development in Chinese grammatical knowledge who require immediate additional interventions.

3.7 A Summary

In order to validate the psychometric properties of the original 172-item Chinese Grammatical Assessment (CGA) and to select good-fit items for developing two normative CGA short tests, namely CGA-A and CGA-B, five phases of research work were conducted with reference to different studies related to the development of language assessment (Wilson, 2001; Efeotor, 2014; Canon & Hubley, 2014; Cannon et al., 2016; and among others). As a comprehensive guide, Efeotor (2014) provides a very good framework that illustrates how Wilson's (2001) "Four Building Blocks" for development of measurements can be applied to the current study. Different validation procedures were incorporated in this study to review the reliability and validity of CGA based on two sets of data, one collected from the project "Profiling Chinese Grammatical Knowledge of Deaf and Hard-of-Hearing Students in HK and China – A Comparative Study" (Tang et al., 2020), another one is a newly collected data in 2022 from the SLCO Programme (Tang et al., 2023) with a big group of DHH students studying with their typically developing peers. As a summary for the research framework of this study, Figure 7 provides a flow chart that describes the methodology and procedures for the different phases of analysis of the study. In the following chapters, we will discuss the results phase by phase.

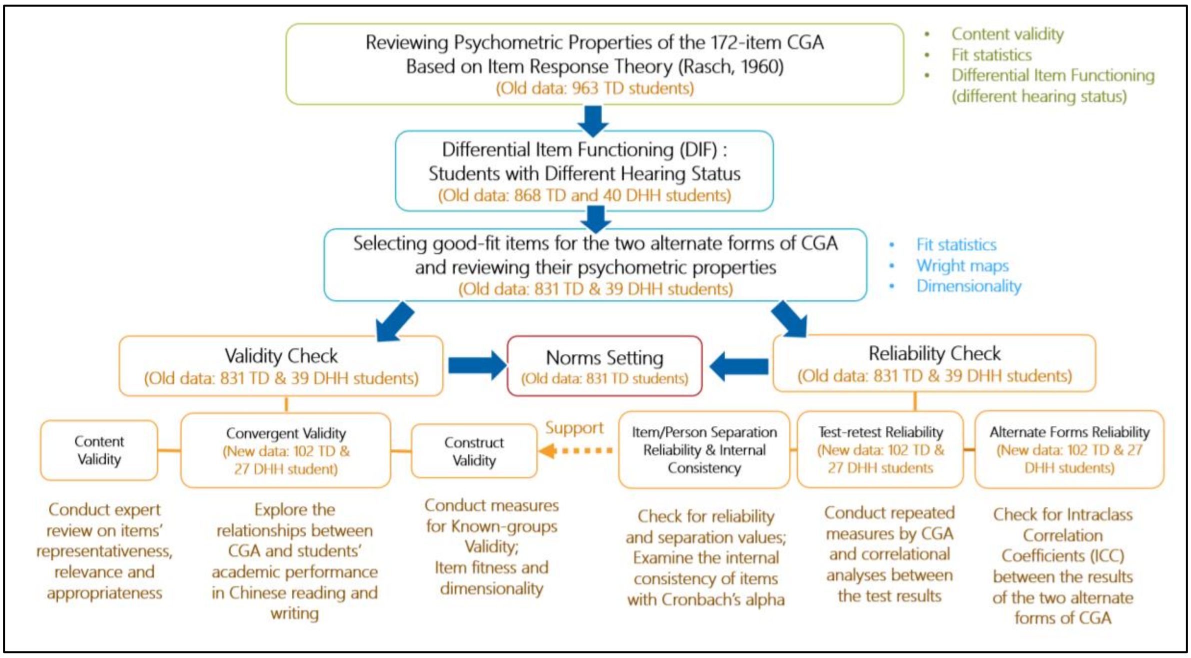

Reviewing Psychometric Properties of the 172-item CGA Based on Item Response Theory (Rasch, 1960)
(Old data: 963 TD students)
Content validity
Fit statistics
Differential Item Functioning (different hearing status)
Differential Item Functioning (DIF) : Students with Different Hearing Status
(Old data: 868 TD and 40 DHH students)
Selecting good-fit items for the two alternate forms of CGA and reviewing their psychometric properties
(Old data: 831 TD & 39 DHH students)
Fit statistics
Wright maps
Dimensionality
Validity Check
(Old data: 831 TD & 39 DHH students)
Norms Setting
(Old data: 831 TD students)
Reliability Check
(Old data: 831 TD & 39 DHH students)
Content Validity
Convergent Validity (New data: 102 TD & 27 DHH student)
Construct Validity
Support
Item/Person Separation Reliability & Internal Consistency
Test-retest Reliability (New data: 102 TD & 27 DHH students
Alternate Forms Reliability (New data: 102 TD & 27 DHH students)
Conduct expert review on items' representativeness, relevance and appropriateness
Explore the relationships between CGA and students' academic performance in Chinese reading and writing
Conduct measures for Known-groups Validity; Item fitness and dimensionality
Check for reliability and separation values; Examine the internal consistency of items with Cronbach's alpha
Conduct repeated measures by CGA and correlational analyses between the test results
Check for Intraclass Correlation Coefficients (ICC) between the results of the two alternate forms of CGA

Figure 7. A summary of the validation procedures for the development of CGA short tests

Chapter 4: Content Validation of CGA

4.1 Results of the Panel Review

“Content validation is a crucial, but often neglected component of good test development” (Cannon & Hubley, 2014, p.768). In order to examine to what extents the item design and operational elements of the Chinese Grammatical Assessment (CGA) are valid for the measurement of TD and DHH students’ grammatical knowledge in written Chinese, 10 subject matter experts (SMEs) were invited to form an expert panel to review not only the grammatical categories and respective items included in the assessment but also its administration and mode of operation such as test instructions, presentation of stimuli, and scoring system. An online platform was then established for the SMEs to review the assessment contents, and at the same time, for an immediate record of the ratings given by the SMEs. In the review, the results in terms of SMEs’ ratings and Content Validity Index (CVI) were used for separate reasons: the ratings were to tap the degrees of endorsement for individual items or areas of CGA development while the CVI was to check the degrees of consensus among the review experts. As mentioned in Section 3.2, the ratings for the representativeness, relevance and appropriateness are based on a 5-point Likert scale. The average ratings of 4 or 5 on one item were considered a positive endorsement of the SMEs whereas the average ratings of 1, 2 and 3 were considered a non-endorsement in the present study. A rating of 4 or above a specific item represents a positive endorsement from one out of ten SMEs, which is equivalent to a CVI of .10. An endorsement on the same item by two SMEs, the value of CVI is equivalent to .20. As a general practice, the highest CVI value is 1.0. Following the suggestion in Lynn (1986), a CVI of .80 (i.e., an endorsement by 8 SMEs) was regarded as significant evidence to justify the content validity of a specific test item. In the following discussions, besides reporting the ratings and CVIs of the review, the SMEs’ suggestions and comments will also be discussed.

The results of the panel review conducted for the 172-item CGA will be discussed according to the following four major areas of investigations. Individual items or operational elements with an average rating <4.0 or a CVI <.80 were flagged for further investigation.

a) Operational elements of the assessment.

b) Representativeness of the grammatical categories involved in CGA.

c) Relevance and appropriateness of the design of individual items.

d) Overall design of CGA.

4.2 Operational Elements of CGA

The first part of expert review is on the mode of operation and administration of the assessment including the mode of operation of the assessment, the instructions to students, the testing procedures, and the mode of responses, etc.

Table 13. Results of Panel Review on the Operational Elements of CGA

Operational Elements of CGA		CVI
1.	Operating as a web-based online assessment.	.90
2.	Displaying items randomly by the computer - every time in a different order.	1.0
3.	Students can change their answers before their submission.	1.0
4.	Using an animated video to explain how to answer the different types of questions.	1.0
5.	The contents and the illustration of the video.	.80
6.	Doing trial items before doing the test items.	1.0
7.	Receiving a vocabulary test before doing CGA.	.90
8.	The number of words in the vocabulary test.	.60[a]

Note. CGA = Chinese Grammatical Assessment; CVI = Content Validity Index.
[a] Operation elements with an average CVI <.80.

As indicated in Table 13, CGA's overall operation was highly endorsed by the SMEs with an average CVI of .90 for the eight questions. Most of the questions regarding the different operational elements received very good ratings, with a CVI of either .90 or 1.0, only that the ratings of Question 8 are relatively low, with an average CVI of .60. The major concern was the size of the vocabulary pre-test for CGA. In the following sections, the results from the expert review will be discussed with an incorporation of the open comments from the SMEs.

4.2.1 Online Mode of Operations

CGA is a receptive test, requiring no writing or typing. An online assessment platform has been developed for CGA so that students can simply respond to the questions by clicking the mouse of their computers or touching the screen of the tablets. All 10 SMEs endorsed the online mode of operation for CGA (CVI=.90), but they also alerted the test operator to be aware of the possible problems that might happen in an online assessment. The following issues were the major concerns raised by the SMEs in their open comments:

i) whether students can technically manage the online testing procedures.

ii) whether guidance and instant help would be available when the students are facing difficulties during the process.

iii) how the test invigilators can ensure that students are doing the test with good attention.

iv) whether a stable network could be ensured during the test.

v) how network disconnection would be handled.

vi) whether environmental disturbance that probably distracts students' attention can be avoided.

vii) whether a "QUERY" button can be added on the test platform so that students can simply press the button for help whenever necessary. The record would also be helpful to review if the students' responses are reliable.

As an online platform, CGA allows flexibility for the students to change their answers easily. This operational element was endorsed by all SMEs (CVI=1.0). What SMEs were concerned about was whether the students were aware of the function and knew how to seek help from invigilators when they faced any difficulties during the assessment. As agreed with the SMEs, this function was difficult to explain clearly through the animated video. Currently, there were trained invigilators attending the assessment sessions and the students were reminded to raise their concerns whenever necessary.

4.2.2 Randomized Presentation of Items

All SMEs agreed that the items should be displayed in a randomized order so that the sequence of item presentation could be changed every time a student attends CGA (CVI=1.0). Considering there were a large number of items included in CGA and the assessment targeted a wide range of students, SME08 suggested developing a computerized system that can randomly select a set number of test items for a student so that students do not need to answer all the 172 questions. This comment leads to two issues: whether the test can be shorter, and the items can be changed every time. In fact, the proposal leads to a discussion about the possibility of developing CAS into a Computerized Adaptive Testing (CAT) (Meijer & Nering, 1999), which can select items automatically from the system and provide optimal test items for individuals according to their performance.

4.2.3 Trial Items and Vocabulary Pre-test

The availability of trial items presented before the testing items was endorsed by all SMEs (CVI=1.0). The arrangement is to help students familiarize themselves with the different types of questions in CGA before they answer the testing items. No special comments received from the SMEs except SME07 who suggested providing an automatic reminder or guidance to the students if they incorrectly answer the trial questions which are supposed to be very simple to primary school students.

Regarding the arrangement of the vocabulary pre-test prior to the main test of CGA, 9 out of 10 SMEs endorsed the arrangement (CVI=.90). The major concern raised by the SMEs was how many vocabulary items should be included. The average CVI is .60. Whether the number of vocabularies in the pre-test is appropriate could not get a straightforward endorsement. though no one SME rated this operational element below 4 (fairly appropriate). In fact, different SMEs had quite different opinions about the size of the vocabulary test. While one SME proposed to include more items in the vocabulary test to cover a wider scope of vocabulary, another SME worried that too many items for the vocabulary pre-test might add too much workload to the students. SME08 doubted if the vocabulary pre-test should be considered obligatory in practice. He added that:

"I think it is appropriate from the point of view of test validity, but considering its practical applications and design, do all students need to have full mastery of all the vocabulary before they can do the test? Should students be excluded from the assessment if they failed the vocabulary pre-test? Or are they still accepted to do the assessment only that their results will be analyzed in a different way?"

In sum, all SMEs welcomed the arrangement of the vocabulary pre-test but there was no consensus from the SMEs because of different reasons. Therefore, no change in the vocabulary test was made.

4.2.4 Animated Video Instructions

Instead of giving verbal instructions to the students for CGA, an animated video is produced to demonstrate how the different types of questions should be answered on the online platform. Using animated video aims to provide equally accessible instructions to both TD and DHH students. No voice-over is available in the video so that DHH students would not be disadvantaged by their hearing difficulties.

All SMEs agreed to the use of the animated video to demonstrate how the different types of questions should be answered (CVI=1.0) though one SME felt not getting used to a video with no sound. Most of the SMEs supported that the contents of the video are appropriate (CVI=.80). Two SMEs gave a rating of 3 in this question, one remarked that the video is a bit long for the students to remember, and the other commented that the digits and emoji representing the procedures disappeared too fast, and the size of the digits were too large and not aligned well with the stimuli, making the video hard for the kids to understand. In sum, the concern is mainly the font size in the video and the pace of the video. One SME suggested having different videos for different task types. However, as all test items, no matter in which types of questions or task types, are all randomly presented in each test, it would be better to use one video for all task types and play it before doing the first item.

4.3 Representativeness and Relevance of CGA

As discussed above, the mode of operations and administration of CGA is basically endorsed by the SMEs though there were some concerns about the vocabulary pre-test. To ensure satisfactory content validity of the assessment, the expert review on the representativeness, relevance and appropriateness of the items and contents of the assessment is essential (Hubley

& Palepu, 2007). The results of their content validity review are discussed in the following sections.

4.3.1 Selection of the Grammatical Categories

Representativeness, in this study, is defined as the extent to which the grammatical categories tested in CGA is reflecting the Chinese grammatical knowledge of primary school students. The 18 grammatical categories were rated by the SMEs individually regarding their representativeness as a Chinese grammatical assessment for primary school students in a 5-point Likert scale (see Section 3.2.2). A low rating on a grammatical category like 1-3 suggests that the grammatical category is not a representative Chinese grammatical knowledge for primary school students. In contrast, a high rating like 4-5 suggests that the grammatical category tested in the assessment is a highly representative grammatical knowledge in written Chinese. All school students are expected to acquire these groups of grammatical knowledge during their primary education.

According to the results summarized in Table 14, the SMEs have positively endorsed the representativeness of the 18 grammatical categories. The average CVI is .90 and the average rating given by the SMEs is 4.49 out of 5.00. The 18 grammatical categories selected for CGA are considered as highly representative of the Chinese grammatical knowledge for primary school students. Among the 18 categories, three categories, namely Cleft Sentence, Question Particles and Binding, got endorsement from only 7 SMEs. Their average ratings on representativeness were relatively lower, with a score of 3.90 for Cleft Sentences, and 4.00 for both Question Particles and Binding. Not all SMEs had given their explanation in the questionnaires. One SME remarked that the Cleft Sentence was a bit hard to comprehend according to the sample sentence *小明是後天參加圍棋比賽* ‘It is the day after tomorrow that

Xiao Ming will participate in the go game'. In view that there were too many categories related to questions, including "Questions", "Question Words" and "Question Particles", one SME suggested to replace the category "Question Particles" with "Sentence Final Particles" so that both interrogative or declarative sentences could be included. Indeed, "Question Particles" is a subset of "Sentence Final Particles", playing a special role in Chinese linguistics (Huang, Li, & Li, 2009).

Table 14. Representativeness of the Grammatical Categories Tested in CGA

Grammatical Categories		CVI	Average Ratings
S01	*ba*-constructions	1.0	4.80
S02	Passives	1.0	4.70
S03	Binding	.70[a]	4.00
S04	Relative clause	1.0	4.60
S05	Comparatives	1.0	4.90
S06	Quantification	.90	4.40
S07	Double-object construction	.90	4.50
S08	Locative existential	.90	4.50
S09	Control	.90	4.40
S10	Cleft sentences	.70[a]	3.90
S11	Question	.90	4.20
S12	Morpheme distinction	1.0	4.60
S13	Negation	.90	4.70
S14	Preposition	1.0	4.60
S15	Localizer	.90	4.70
S16	Aspect	.90	4.70
S17	Question words	.90	4.60
S18	Question particles	.70[a]	4.00
	Average score:	.90	4.49

Note. CGA = Chinese Grammatical Assessment; CVI = Content Validity Index.
[a] Endorsement of less than 80%.

Regarding the category "Binding", three SMEs gave their comments. They explained their concern that there might be more than one interpretation for the sentence *小明的哥哥在畫他*

‘Siu Ming’s brother is drawing him’, which might create ambiguity in students’ understanding of the sentence’s meaning. The pronoun 他 ‘him’ in the sentence can represent anyone except Siu Ming’s brother. This phenomenon is explained by the binding theory, the pronoun 他 ‘him’ in the sentence is free in their governing category (Huang, Li, & Li, 2009).

The SMEs had given some suggestions to include more grammatical categories in CGA such as modal words, demonstratives, classifiers, aspect makers and different types of connectives, for example, 和 ‘and’, 或 ‘or’, 而且 ‘also’, 雖然 ‘although’, 但是 ‘but’, 而且 ‘also’, 因爲 ‘because’, 所以 ‘therefore’, etc. These are relevant categories of the school curriculum, which can be considered for inclusion in the assessment in future.

4.3.2 Relevance and Appropriateness of the Items

All 172 items of CGA were reviewed individually by the SMEs and rated for their appropriateness in terms of the design and their relevance as an item of a Chinese grammatical assessment for primary school students in Hong Kong. Written comments from SMEs were also invited for each item. Their feedback was seriously considered during the process of item selection or item enhancement for the two short tests.

As a summary, the average rating for the items’ appropriateness and relevance was 4.64 (ratings ranged from 4.00-5.00) and 4.74 (ratings ranged from 4.20-5.00) respectively. Their average CVIs were .91 and .95 respectively. With an average >.90, according to Hubley and Palepu (2007), CGA was positively endorsed by all the ten SMEs. Most of the items were positively endorsed by the SMEs regarding their relevance to an assessment for the Chinese grammatical knowledge (CVI=1.0). The 10 (out of 172) items with their CVIs <.80 are flagged for scrutiny. These items belong to four grammatical categories including Comparatives (6 items), Binding (2 items), Localizer (1 item) and Relative Clause (1 item). No items in CGA were considered irrelevant by the SMEs, only some items were considered "fairly appropriate", with the lowest ratings of 4.00 out of 5.00. The major comments of SMEs on these items were summarized below:

i) Comparatives

The SMEs endorsed all items with basic comparatives, however, for the six items with negated comparatives *不比* 'not-compare', they only gave a fair endorsement on their appropriateness. One SME was concerned about the design of the pictures that might cause interference with students' responses. The other two SMEs were concerned about the comprehension difficulties arising from the incorporation of negation in the comparative constructions. They found the construction *X 不比 Y 高* 'X is not taller than Y' a bit "ambiguous" to the readers. It might be caused by the two possible interpretations of this kind of non-strict comparatives (Nouwen, 2008) including: i) X is shorter than Y (interval reading); and ii) X is as tall as Y (equality reading). Primary school students may find it hard to accept both readings. However, as an assessment to check for students' understanding about the sentence structure according to its morpho-syntax. It is worth checking if the students can accept the two readings represented by the structure.

ii) Binding

For the two items under the category of Binding, the endorsement from SMEs was "fair". One SME commented mainly on the choice of action verb *打* 'hit' which is aggressive in nature and the design of the pictures that show how the "grandmother" is hit by herself or somebody.

The SME would prefer using more positive actions to construct the items. Basically, the concern was the negative connotation delivered through the pictures, rather than the design of the design of the items.

iii) Localizer

Items regarding Localizer are tested by the truth value judgement task in CGA. Though other SMEs had no special concern about the task chosen for these items, one SME proposed that it would be more reliable to test the concept of Localizer by a picture selection task rather than a truth value judgment task. The SME's suggestion would focus more on the meaning of the different localizers such as *上* 'up' and *裏* 'inside', in which, whether students can get the correct answers only depends on their knowledge about the lexical meaning of the localizer, not the syntax.

The current test items would focus more on the syntactic form of the constructions, with or without localizers. For example, the presence of localizer is obligatory according to native speakers of Mandarin, but it is not always obligatory. In another word, the focus is to check the metalinguistics awareness of the students in both Cantonese and written Chinese (Lau et al., 2019).

iv) Relative Clauses

There is one relative clause item not fully endorsed by the SMEs. Their concern was that the design of the picture cannot clearly represent the sentence. After thorough scrutiny, the category of relative clause and the one concerned item is worth keeping it in the two short tests at this stage, only that more frequent review may be required

v) Other comments

Besides the comments given for the above grammatical categories, there are some other suggestions and recommendations for some items. As a summary, the areas of recommendations from individual SMEs including the following:

i. Some pictures can be further modified for better representation of the stimuli, the distractors or the answers, such as some items in the categories Binding and Relative Clause.

ii. There were too many similar items for some grammatical categories such as Quantification, Questions, Morpheme Distinction, and Relative Clause. Some items can be excluded from the two short tests.

iii. Some items are difficult for primary school students and their meanings are ambiguous and difficult to grasp clearly.

All the recommendations given by the SMEs were considered when item selection was conducted for the establishment of the two short tests, considering their content validity for the assessment. Some items would be considered modifying in future based on the recommendations from different experts when CGA is further re-structured and modified for another round of norm setting.

4.4 Overall Design of CGA

As showed in the Table 15, five questions were designed for the SMEs to review and comment on the overall design of the assessment. Results showed that they endorsed favourably the title of the test though some of them also provide suggestions for the name. Their suggestions mainly emphasized how the name can be modified to better match its aim and target. SME06 commented that the title of the assessment should include the concept of "Receptive Grammar" to highlight that CGA is a comprehension test. Regarding the title of the assessment, SME02 suggested changing the name to *香港兒童中文語法評估* (*Chinese Grammatical Assessment for Children in Hong Kong*), aimed to highlight that primary school students are the target of CGA. As a long-term development, the name *小學中文語法理解評估 (Receptive Assessment on Chinese Grammatical Knowledge for Primary School Students)* may be a good alternative which reflects both the aim, the skills tested and the target of the assessment.

Table 15. Results of Panel Review on the Overall Design of CGA

Questions		CVI
1.	How appropriate is the title "Chinese Grammatical Assessment (中文語法評估)"?	.80
2.	How appropriate are the overall operations of the assessment?	1.0
3.	Are the selected 18 grammatical categories of CGA having good representativeness in assessing primary school students' grammatical development in written Chinese?	1.0
4.	Are test items of CGA suitable for testing Chinese grammatical knowledge of deaf or hard-of-hearing children?	.90
5.	Are test items of CGA suitable for testing Chinese grammatical knowledge of typically developing children?	.80

Note. CGA = Chinese Grammatical Assessment; CVI = Content Validity Index.
[a] Endorsement of less than 80%.

The overall operations of CGA were considered appropriate according to the SME's ratings (CVI=1.0). The 18 grammatical categories selected for CGA were having good representativeness in assessing primary school students' grammatical development in written Chinese (CVI=1.0). In addition, the SMEs considered that the items used in CGA were suitable items for testing Chinese grammatical knowledge of both TD (CVI=.80) and DHH students (CVI=.90).

4.5 Development of Two Short Tests

Other than the results analyzed above, according to the written comments of the SMEs, some SMEs expressed concern about the large number of items (*N*=172) used in the assessment might overload the children. One SME commented that:

> "*As there are quite a large number of test items included in CGA, a good concentration is demanded for students to analyze the stimuli and the meaning of the pictures. I am worried that the students, especially the junior ones, would not be able to keep their attention or maintain their physical strength to complete the assessment all at once. Students may simply mess around when they do the assessment, which may affect the reliability of the results. Would it be good to give students a break in the middle of the assessment to 'charge them up' before continuing to do the remaining questions?*"

As the SMEs did not know the intension of developing two alternate forms of CGA when they conducted the expert review, they still assumed that the final version of CGA would be a 172-item assessment. To clarify this issue and to collect their views on the development of two CGA short tests, a follow-up question "Regarding your professional work, what do you think if we develop two CGA short tests with norms, each of which includes about 45-50 items based on the long version?" was sent to the SMEs individually for their further comments. Six SMEs replied to the question. They all agreed that it is conducive to develop two CGA short tests based on the original long version with 172 items in CGA. Their reasons for supporting the proposal are summarized as below:

(1) **More efficient:** They believed that a short test with 45-50 items would be much easier for children to complete.

(2) **More reliable results:** They believed that the students would be more attentive, and the results would be more reliable.

(3) **Practically appropriate:** They commented that a shorter version would be more useful for professionals in daily clinical practices. It was time saving that would be able to finish in one training session.

(4) **Tracking students' development:** Having two alternate forms can be used for pre- and post-test comparisons, which help keep track of students' progress and develop their treatment plans.

Even though the SMEs welcomed the idea of developing two CGA short tests for educational and clinical use, one SME reminded that the two tests had to be systematically reviewed on their validity and reliability, and they should have well standardized norms for accurate identification of the needs of both TD and DHH students.

4.6 An Interim Discussion

To review the content validity of the Chinese Grammatical Assessment, an expert panel with 10 members including speech therapists and teachers were set up for the review of the representativeness, appropriateness and relevance of the grammatical categories and respective test items were reviewed extensively according to a questionnaire. Besides the design and operations of the overall assessment, all 18 grammatical categories and 172 test items were reviewed.

The results of the content validation are very positive based on the ratings and the projected content validity index (CVI). The Subject Matter Experts (SMEs) endorsed the design of the assessment including its title, its operation and administration. The selected grammatical categories and the test items were given very high ratings from the SMEs. In sum, the representativeness of the 18 grammatical categories was endorsed with an average CVI of .90. In addition, 171 out of 172 items (99.42%) and 162 out of 172 items (94.19%) received a CVI >.80 regarding the appropriateness and relevance of the items for assessing Chinese grammatical knowledge of primary school students respectively. There were comments given by SMEs, concerning the design of some specific items, the pictures used in the item, or the difficulties of the items for primary school students. The SMEs also provided some useful suggestions for further development of the assessment in future. In the following section, the assessment would be reviewed further based on its psychometric properties as well as its other measures of validity and reliability. Together with the results of the content validation reported in this section and the specific comments given by the SMEs on individual items, two equivalent lists of items would be selected for the development of two CGA short tests for educational and clinical practices.

Chapter 5: Psychometric Review of the Items

The long version of the Chinese Grammatical Assessment (CGA) consists of 172 items for profiling DHH students' grammatical development. With the consent of the Centre for Sign Linguistics and Deaf Studies, The Chinese University of Hong Kong, a dataset was extracted from the database for further analysis of the psychometric properties of the items. As mentioned in Chapter 3, the 963 CGA data from typically developing students and 40 data from DHH subjects were used for the psychometric analysis for CGA. Good-fit items comprising different representative and relevant grammatical knowledge would be selected for the development of the two equivalent lists of items, and eventually the two CGA short tests for the assessment of primary school students' Chinese grammatical knowledge.

Grammatical knowledge in written Chinese is the latent trait that CGA is targeted to assess. As discussed in Chapter 4, a vocabulary pre-test with 32 items was created to check if the students understand the major words used for the items of CGA. As the assessment focuses on students' comprehension of different grammatical constructions, this arrangement is to prevent students' insufficient knowledge of basic vocabulary from confounding the test results. In this study, an accuracy rate of 75% (24 out of 32 words) is used as the cut-off point, which means, the data with a vocabulary pre-test score below 75% would be removed from the current dataset before further analysis. Following this criterion, 56 data with the vocabulary test results < 75% were excluded and thus the remaining 907 data were used as a dataset for the psychometric analysis of the 172-item CGA for the development of the two CGA short tests. No data from DHH subjects were deleted because of the vocabulary pre-test (see Table 16).

The scores of CGA are either "0" and "1". According to the Item Response Theory, the

Dichotomous Rasch Model (Rasch, 1960) was used for the analysis. In this regard, different aspects of statistical reviews were conducted based on fit statistics, person and item reliability, and differential item functioning. Item-level information is more important at this stage to help select the good-fit items from the 172 item-pool. This is why IRT was used instead of the Classical Test Theory.

5.1 Fit Statistics

Fit statistics are important procedures in Rasch analysis as it can give crucial evidence for the psychometric validity of the measurement and the conformity of the data to the predictions of the analysis (Aryadoust, Ng, & Sayama, 2021). Therefore, besides calculating the item difficulty and person ability from the data, fit statistics were generated to see if the data fit the model. The values of fit statistics are close to one when the data fit the model. With reference to Linacre (2002) and Bond and Fox's (2007), the infit or outfit mean-squares (MNSQ) of the persons and items should be within the range of 0.50 and 1.50 for constructive development of the test (Wright & Linacre, 1994) and the selection of best-fit items for the finalized two short versions of CGA.

Fit statistics were performed for the review of students' person ability and then the items' difficulty. MNSQ values for infit or outfit statistics higher than 1.50 imply an underfit while the MNSQ values below 0.50 denote an overfit (Bond & Fox, 2007). The underfit data are less predictable than the model expects, while the overfit data are more predictable than the model expects (Wright & Linacre, 1994). Data with a person outfit MNSQ values out of the acceptable range were removed from the dataset before going through the items' fit statistics.

After performing two rounds of persons' fit statistics, 39 data were excluded. No persons' infit MNSQ, but 39 outfit MNSQ was found to be out of the acceptable range of 0.50-1.50. Therefore, 868 TD and 39DHH data were kept for further item statistics (see Table 16 for a summary of the changes of the dataset), ranged from 103 to 226 data for different grade levels (see Table 17).

Table 16. Change of dataset after vocabulary pre-test and fit statistics for persons' data

	TD data	DHH data
Original dataset of the 172-item CGA	963	40
Data taken away:		
a. Vocabulary pre-test <75%	-56	/
b. Misfit data taken away after Fit statistics	-39	-1
Remaining data for further Rasch analysis:	868	39

Table 17. Number of TD and DHH data of primary school students for a combined analysis

	P1	P2	P3	P4	P5	P6	**Total**
TD students	95	177	219	101	164	112	**868**
DHH students	8	8	7	6	4	7	**40**
Total:	103	185	226	107	168	119	**908**

After fit statistics for persons, the misfit persons' data were removed from the dataset, so item fitness was conducted based on the remaining 868 data as listed in Table 17. Four items had an outfit MNSQ >1.50. Scrutinizing these 4 items, they basically belong to two types of grammatical constructions, including two items from Comparatives (with the outfit MNSQ of 1.60 and 1.72), and two items from Localizer (with the outfit MNSQ values of 1.65 and 1.97). The section below will be discussing about the specific linguistic properties of the two structures and the possible reasons for their identification as outfit items, which would be kept in the item pool at this stage and further reviewed for their psychometric properties during the item selection process.

5.1.1 Considerations for Two Outfit Grammatical Sub-categories

Among all items from the 18 categories in Chinese grammatical knowledge, there were two items in each alternate forms identified as outfit items in the psychometric review and finally deleted from the two lists of CGA. They belong to the two grammatical categories, namely Comparatives (cnbbPM03 and cnbbPM04) and Localizer (lonlGJ02 and lonlGJ03). These items represent different linguistic properties of Chinese grammar. In the following sections, we will describe these two specific grammatical categories and explain why they were eventually excluded in the two equivalent lists.

5.1.1.1 Comparatives (with Negation)

One of the two relatively difficult grammatical sub-categories in CGA belongs to "Comparatives (with Negation)". The sub-category is testing students' understanding of the concept 不比 'not-compare', which is the combined use of the negator 不 'not' and the comparative marker 比 'compare'. As a Truth Value Judgement question, students are required to judge if the meaning of the sentence matched with the stimuli, in the form of a picture (see sentences (10) and (11) as examples of the sub-category).

(10) 我 爸爸 比 哥哥 高 (comparative with no negator)

My father compare brother tall

'My father is taller than my brother.'

(11) 我 爸爸 不 比 哥哥 高 (comparative with negator)

My father not-compare brother tall

'My father is not taller than my brother.'

For the sentence (10), there is no negator attached to the comparative marker, the meaning is more straight forward with only one interpretation, that is, "Height of father > Height of brother". However, when considering the example (11) above, as a kind of non-strict comparatives, there are two acceptable interpretations (Nouwen, 2008): (i) "Height of father < Height of brother" (interval reading); & (ii) "Height of father = Height of brother" (equality reading). According to the results, students might find difficult to accept the equality reading for the sentence in written Chinese.

In Cantonese, the equivalent comparative sentences of (10) and (11) are structured like the sentences (12) and (13).

(12) 我 爸爸 高 過 哥哥 (comparative with no negator

My father tall more-than brother in Cantonese)

'My father is taller than my brother.'

(13) 我 爸爸 高 唔 過 哥哥 (comparative with negator

My father tall not more-than brother in Cantonese)

'My father is not taller than my brother.'

The word order for this negated comparative in Cantonese is quite different from that in written Chinese. The word order in written Chinese is *A 不比 B 高* ‘A not compare B tall’, the adjective *高* ‘tall’ in written Chinese is in the sentence final position. However, for the equivalent sentence in Cantonese, the word order is *A 高唔過 B* ‘A tall not more-than B’. The sentence structure is very from that in written Chinese and the adjective *高* ‘tall’ is situated in the sentence-middle position.

There is another negated comparative sentence in Cantonese, which has a word order *A 唔夠 B 高* ‘A not-enough B tall’ (see the sentence (14) below), which is very similar to the word worder in written Chinese (see the sentence (15) below), but the interpretations of these two structures are different. There is only one interpretation accepted for the Cantonese sentence, which is “Height of father < Height of brother”. Unlike the Chinese sentence, the same-height interpretation, that is, “Height of father = Height of brother” is not acceptable in this sentence.

(14) 我 爸爸 唔夠 哥哥 高 (comparative with negator
My father not-enough brother tall in Cantonese)
‘My father is shorter than my brother.’

(15) 我 爸爸 不 比 哥哥 高 (comparative with negator
My father not-compare brother tall in written Chinese)
‘My father is not taller than my brother.’

We cannot find any research evidence to verify our intuition or hypothesis. It is not clear whether the phenomenon can be explained by the linguistic differences between Cantonese and written Chinese, but this would be an important research topic if we want to understand the specific difficulties facing Cantonese-speaking students in their development of early literacy in written Chinese. The items required relatively high level meta-linguistic awareness before they can identify the specific differences between the grammar Cantonese and written Chinese. It may be too difficult for the target group of students.

5.1.1.2 Localizers

For the other four items with outfit measures >1.50, they belong to the category “Localizer”. The items were found to be relatively more difficult among all items in CGA. These items are designed to see if the students are aware of the obligatory status of “localizer” in the location expressions of a sentence (see examples (16) and (17)). For example, in written Chinese, the sentence (16) is ungrammatical because of the absence of localizer such as *裏面* ‘inside’.

(16) *媽媽 在 餐廳 (Locative expression in written Chinese)
Mother in restaurant
(17) 媽媽 在 餐廳 裏面 (Locative expression in written Chinese)
Mother in restaurant inside
‘Mum is in the restaurant.’

The absence of localizer *裏面* ‘inside’ in (17) is not acceptable in written Chinese. In contrast, it is acceptable to express the same idea in Cantonese with no specific localizer in the sentence like (18). The localizer *裏面* ‘inside’ in (18) is optional in Cantonese.

(18) 媽媽 喺 餐廳 (裏面) (Locative expression in Cantonese)

Mother at restaurant (inside)

‘Mum is in the restaurant.’

A more general localizer marker 度 ‘there’ can also be used in Cantonese, in apposition to a noun phrase to indicate a specific place or location the subject is located (Matthews & Yip, 2011) (see example (19)). CL in the sentence represents “classifiers” which are used to express quantities of mass nouns (Matthews & Yip, 2011, p.72).

(19) 媽媽 坐 喺 張 凳 度 (Locative expression in Cantonese)

Mother sit at CL chair there

‘Mum is sitting on the chair.’

As explained above, localizer is not obligatory in the locative sentences in Cantonese, it may be quite difficult for Cantonese-speaking primary school children to identify this linguistic difference and realize that the status of localizer is different in written Chinese. Therefore, like the outfit items (under the category “Comparative”) we discussed above, even high ability children might believe that the localizer could be dropped and thus did not reject the locative sentence with no localizer. This may also be a reason to explain why these items did not fit well with the model.

In view that the outfit MNSQ values of the four items in the two grammatical sub-categories, namely Comparative and Localizer were all >1.5. They were all considered not the good-fit items for CGA. In addition, the content validation results for the four items were dissatisfactory, having a relatively low CVI values (one item with CVI=.60, two items with CVIs=.70, and one item with CVI=.80), these four items and the two respective sub-categories were thus excluded from the item pool. In this regard, for each alternate form of CGA, there were 46 items selected from 46 grammatical sub-categories.

5.2 Person and Item Reliability

As stated in the last section, after conducting fit statistics, 868 TD and DHH data were left for further analysis. Rasch analyses were conducted for the TD subjects only, and the "TD+DHH" subjects separately to ensure that the model fits well with and without DHH students. Results indicated that the separation reliability of CGA was good for both conditions: person and item reliability are .96 and .99 respectively and the separation index for person and item statistics are both >2 (see the results summarized in Table 18 for both conditions), reflecting that the proportion of true variance (or adjusted variance in Rasch term) relative to the error variance is high (Guilford, 1965).

The results for the two conditions were very similar according to the descriptive statistics listed in Table 18. The means for person ability were 54.65 (SD=8.20) and 54.66 (SD=8.21) for the TD and the combined TD+DHH condition respectively. The means for item difficulty were 46.53 (SD=5.62) and 46.53 (SD=5.65) for the TD and TD+DHH condition respectively. The mean person ability is 2.58 and 2.56 logits higher than their mean item difficulty in the TD and TD+DHH conditions respectively. This implies that the overall items included in CGA were relatively easy for the assessed participants. It would be better to include more difficult items in the assessment for the high-ability group.

Table 18. Results of Rasch analysis of the 868 TD data and a combined group of TD and DHH students (N=907)

	868 TD data		868TD + 39DHH data	
	Person (*N*=868)	Item (*N*=172)	Person (*N*=907)	Item (*N*=172)
Mean	54.65	46.53	54.66	46.53
Standard Deviation	8.20	5.62	8.21	5.65
Separation	5.07	8.63	5.07	8.87
Reliability	.96	.99	.96	.99

Remark: TD=typically development students; DHH=deaf and hard-of-hearing students

To further evaluate the psychometric properties of the items, Differential Item Functioning (DIF) of the items was conducted to see if there were items biased toward either TD or DHH subjects. A combined dataset (N=907 data) with 868 TD and 39 DHH data was used for the analysis of Differential Item Functioning (DIF). The results of DIF can provide significant information that facilitates the item selection process for the alternate forms of CGA.

5.3 Differential Item Functioning (DIF)

"A fair test should be no bias amongst any subgroup" (Efeotor, 2014, p.218). A standard investigation of IRT was conducted to see if the items used to assess students' Chinese grammatical knowledge were appropriate for both typically developing (TD) and DHH students. Based on Linacre (2012), the test for Differential Item Functioning (DIF) is to check if there are items disadvantaged to either group of participants.

As a field testing, the size of the dichotomous dataset should better be >1000 for each sub-group to get a robust DIF analysis (Scott et al., 2009). As recommended by Linacre (2012), when the sample size is small, it is still worth performing DIF in terms of a trial test, to help identify possible group differences and individual items that may need to be reviewed before including in the assessment. The information projected from the analysis should still be a good reference during the process of item selection and refinement. In addition, the results DIF are good indicators that help to bring up possible precautions or special attention we need to be aware of when the test is applied to the target groups (Efeotor, 2014).

Following the criteria proposed by Zwick, Thayer, and Lewis (1999), for items with a significant *p*-value <.05 in the Mental-Hsenszel Chi-square test and the absolute DIF Contrast (|DIF|) > 2 logits, they would be flagged for further investigations. To perform DIF testing regarding students with different hearing status, the 39 data from DHH students studying in two mainstream primary schools were combined to the 868 data from TD students before conducting the analysis.

The results showed that a total of 20 items were found to have *p* <.05 in the DIF measures. With reference to the values of the DIF Contrast, nine items were found to be significantly less favourable to the DHH students and 11 items were significantly less favourable to the TD students (see Table 19). Scrutinizing the items with significant DIF contrast, the major findings are summarized in the following sections. Final inclusion or exclusion of items in the two short tests were considered together with other reviews, including the fit statistics, the ratings and CVI of individual items, the design of the items. Relatively good items would be kept as far as possible to maintain a wider spectrum of grammatical constructions in the two CGA short tests. For each sub-category of grammatical knowledge, two items were selected, one item for one short test. Therefore, some items were excluded simply because there were sufficient items for the two tests.

Table 19. Results of the Differential Item Functioning (DIF) analysis for the 172 items of CGA

Grammatical Knowledge		DIF Measures		DIF Contrast	Mantel-Haenszel		Included or Excluded[b]
Categories	Item Code[a]	TD	DHH		Chi-square	*p*-value	
Items less favourable to DHH subjects:							
Aspect	aspfTV05	52.50	58.58	-6.08	5.34	.0208	excluded
Ba-construction	bacoGJ04	37.21	45.89	-8.68	8.33	.0039	included
Passives	beixTV01	44.91	51.96	-7.05	5.62	.0177	excluded
Comparatives	cnbbPM02	45.02	50.04	-5.01	5.99	.0144	excluded
Negation	negmFB03	40.99	49.04	-8.05	8.10	.0044	included
Question particles	qpmaGJ02	55.06	63.79	-8.73	7.39	.0066	excluded
Question particles	qpmaGJ03	54.33	62.67	-8.35	6.83	.0089	included
Question particles	qpmaGJ04	54.33	63.79	-9.46	7.04	.0080	included
Relative clause	rcosPS02	50.01	58.58	-8.58	10.10	.0015	included
Items less favourable to TD subjects:							
Aspect	aspfTV07	47.66	29.76	17.9	8.26	.0041	included
Aspect	aspfTV08	59.19	52.91	6.28	3.84	.0500	included
Aspect	aspgTV02	38.38	15.23	23.15	5.17	.0230	excluded
Ba-construction	baxxTV01	44.97	29.76	15.21	6.97	.0083	excluded
Ba-construction	baxxTV02	38.85	24.31	14.54	4.87	.0273	excluded
Ba-construction	baxxTV04	51.13	45.89	5.24	4.48	.0343	included
Binding	bnpnPS03	51.40	42.26	9.13	8.50	.0035	included
Control	ctocPS03	50.95	40.87	10.08	9.36	.0022	included
Negation	nqqnPS02	43.29	29.76	13.53	5.48	.0193	included
Relative clause	rcsoPS02	40.17	33.10	7.07	3.93	.0473	excluded
Relative clause	rcsoPS03	42.46	29.76	12.70	5.06	.0244	excluded

[a] Specific item codes were designed for individual items according to their grammatical sub-categories, task types. and item number under that category.

[b] Final inclusion or exclusion of items in the two short tests were considered together with other reviews, including the fit statistics, the ratings and CVI of individual items, the design of the items. Relatively good items would be kept as far as possible to maintain a wider spectrum of grammatical constructions in the two short tests.

5.3.1 Results of DIF Analysis

The results of the DIF analysis showed that some specific grammatical categories were less favourable to the TD students, and some were less favourable to the DHH students. Specifically, "Question Particles" (3 items) was found to be less favourable to the DHH students, whereas "Aspect" (3 items), "*Ba*-constructions" (3 items), and "Relative Clause" (2 items) were less favourable to the TD group. There were also some grammatical categories, including

“Negation”, “Binding” or “Comparatives” comprised of one item less favourable to the TD students, but another item less favourable to the DHH students. From our observation, there were no specific patterns or characteristics of these items related to students’ hearing status and would cause any additional disadvantage to their understanding of the stimuli or answers.

For all items listed in Table 19, almost all of them had an average rating ≥4.5 by the 10 SMEs, and their CVIs were either .90 or 1.0, except two items under the grammatical categories of: i) Comparatives (with an item code of cnbbPM02) and ii) Relative Clause (with an item code of rcsoPS02). Their average ratings and CVIs regarding their relevance in the assessment were high (average ratings=4.6; CVIs=.90 for the two items), but their average ratings and CVIs regarding their appropriateness of item design were both lower than the other items (average ratings=4.3; CVIs= .70 for the two items). For different items we had different considerations, but these two items were excluded from the two short tests because they were identified with some item design issues by the SMEs.

5.3.2 Items Less Favourable to the TD Students

Among the 18 grammatical categories, “Aspect” (3 items), “*Ba*-constructions” (3 items), and “Relative Clause” (2 items) were the three grammatical categories that found to have more than one item with p<.05 in the DIF analysis. These items were statistically less favourable or more difficult to the typically developing students when compared to the results of the DHH students. So far, to the knowledge of the author, there is no related research studies investigated on the acquisition of these structures in typically developing Cantonese-speaking students. It is not easy to give a simple conclusion to explain the phenomenon, but it is worth investigating further in the future to understand why these items would be more favourable to the DHH students. Would it be the influence of Cantonese grammar, which may have a greater impact on TD students' acquisition in written Chinese? No matter what the reasons are, these items were of lower priority in item selection for the two CGA short tests.

5.3.3 Items Less Favourable to the DHH Students

There were also items found to be more difficult for the DHH students based on the DIF analysis. As mentioned in Section 5.3.1, three items under the category "Question Particles" were identified as significantly less favourable to the DHH students. Literature reviews were conducted but no specific evidence could be found to clearly give an explanation for the result. In fact, question particles or sentence final particles are important morpho-syntactic knowledge in Cantonese (Matthews & Yip, 2011) and in Mandarin Chinese (Yip & Rimmington, 2004). These particles serve important communicative functions such as defining the types of speech-acts (a question or a request) or expressing specific emotions (Matthews & Yip, 2011).

According to anecdotal observation, these particles, as some special functional categories in Chinese, are auditorily unstressed in daily speech acts. It is always difficult for students with significant hearing loss to perceive them and understand their semantic meanings and specific function in the language. As in Mandarin Chinese, question words like *嗎 (maa1)* and *呢 (ne1)* are serving different functions. The former serves more as a query and the latter a rhetorical question (Yip & Rimmington, 2004). These subtle differences projected by these two question particles may create additional difficulties for DHH students (de Villiers, de Villiers and Hoban, 1994). In view that these items were all having very positive ratings on their relevance (ratings=4.6 and CVI=.90 for all items) and appropriateness (ratings=4.5 and CVI=.90 for all items) in the panel review, they were all kept in the assessment for the development of the two alternate forms of CGA. How difficult the question particles in written Chinese for the DHH students are is still a question yet to be explored. This study may probably provide some insights for educators or speech and language therapists to understand more specifically the needs of DHH students in developing question particles or sentence final particles in written Chinese.

5.4 An Interim Discussion

After conducting content validation for the items of CGA, the psychometric properties of the items based on Rasch analysis was conducted to help select good-fit items for the two CGA short tests. Fit statistics were performed to evaluate all the 172 items in the original item pool. The results indicated that there were only 4 items in two grammatical sub-categories that did not fit well with the model, and their content validation results were not positive, these items were thus excluded from the two alternated forms.

Analysis based on Differential Item Functioning (DIF) further reviewed the items to ensure that there were no items biased to either the TD or the DHH students. Twenty items were found with statistical significance in Mental-Hsenszel Chi-square test, and 9 items were finally excluded from the two alternate forms (see Table 19) according to the results from different analyses. Once the two alternate forms were finalized, the norms of the two respective short tests for Chinese grammatical knowledge were also established according to the data from TD students. In the following chapters, some psychometric reviews of the two alternate forms will be reported according to the Rasch analysis and measures for their different aspects of validity and reliability.

Chapter 6: Finalizing the Two Alternate Forms of CGA

6.1 Item Selection and Validation

The development of two CGA short tests is of multiple advantages in different educational and clinical applications: i) the students can complete the assessment with less time and better attention; ii) the assessment can be conducted more efficiently; iii) the two equivalent CGA short tests can be used to track students' development interchangeably; and iv) it allows more rooms for further inclusion of representative items in CGA.

To ensure that the items of the two alternate forms, namely CGA-A and CGA-B, were valid and reliable, a series of procedures for item selection and psychometric review were conducted as follows:

i) Reviewing content validity of the items selected for the two alternate forms of CGA.

ii) Forming two alternate forms of CGA with comparable item difficulties.

iii) Checking for the alternate forms reliability between the two lists of items.

iv) Reviewing basic psychometric properties of the two forms.

v) Confirming validity and reliability of the two forms with newly collected data.

vi) Norm setting for the two alternate forms, which will eventually be developed as two CGA short tests.

6.1.1 Determining Grammatical Categories and Corresponding Items

In view that the selection of the 18 grammatical categories were all endorsed by the SMEs of the expert panel in terms of their representativeness for the assessment of grammatical knowledge in written Chinese of primary school students (average CVI=.90) (with reference to Hubley & Palepu, 2007) (see Section 4.3.1). The test items were also supported by the expert panel as appropriate (CVI= .91) and relevant (CVI= .95) items for the development of a Chinese grammatical assessment for primary school students. With the endorsement of the SMEs, two preliminary forms of CGA, each comprised of 46 items, one item selected from one of the 46 grammatical sub-categories. Therefore, all 46 grammatical sub-categories were tested in the two short tests. In this regard, a wider coverage of representative grammatical knowledge of written Chinese in the two short tests could be ensured.

After the previous validation processes, two 46-item CGA short forms were developed. The items were selected based on the results of their content validity, fit statistics, DIF results, and their item difficulty, aiming to match the two CGA short forms' overall item difficulty. After item selection, to ensure that the two alternate forms both fitted well with the construct, fit statistics was conducted to review the psychometric properties of the two short forms. As mentioned before, in addition to the outfit mean square values (outfit MNSQ), the outfit z-standardized values (ZSTD) and the Point Measure Correlation (PTMEA-CORR) would also be reviewed.

6.1.2 Establishing Two Alternate Forms with Comparable Item Difficulties

Item fitness to the model is an important factor affecting the validity and reliability of the measurement (Efeotor, 2014). As the two alternate forms would be developed as two short tests for Chinese grammatical knowledge, the selected items for the two forms should be well-fit to the test model. A series of Rasch analyses were conducted with the two forms. Before checking item fitness, person fitness was reviewed beforehand according to the criteria of infit MNSQ and outfit MNSQ >0.5 and <1.5. The item difficulty measures of the items selected for the two forms were tested with their alternate forms reliability to ensure that the results of the two short tests would be comparable with each other.

To recapitulate the development process, the original dataset collected for the 172-item CGA included data from 963 TD students. After screening from the vocabulary pre-test, and a few rounds of fit statistics on the person ability measures, 95 data were deleted, leaving 868 data for the initial development of the two alternate forms of CGA. When the two preliminary alternate forms were confirmed, further Rasch analyses were conducted to ensure that the person data fell within the range of 0.5-1.5 in their infit and outfit MNSQ. During this process, 37 data were further removed from the dataset, leaving 831 TD data for further validation of the two alternate forms. Table 20 summarizes the changes of the dataset through the above-mentioned review process and the data's distributions in terms of the six grade levels. P1 data were deleted the most because of their failure to achieve the 75% standard in the vocabulary pre-test. With this finalized dataset, item fitness was reviewed again.

Table 20. Number of data from typically developing (TD) students for the review of the two alternate forms of CGA.

TD Students	P1	P2	P3	P4	P5	P6	**Total**
Dataset for the review of the two alternate forms	95	177	219	101	164	112	**868**
Data Clearance after Fit Statistics	11	7	3	1	5	10	**37**
Finalized Dataset	84	170	216	100	159	102	**831**

6.1.2.1 A Joint Analysis for the Two Alternate Forms

As mentioned in Section 5.1.1., two sub-categories were excluded from the original 48 grammatical sub-categories, leaving 46 sub-categories in CGA. By selecting one item from one sub-category, two 46-item alternate forms were developed for further review before confirming them as the two finalized CGA short tests. For better comparison between the logits of the two forms, a joint analysis was then conducted to estimate the item difficulty of the items included in the two alternate forms. Table 21 shows the measures of the 92 items included in the two alternate forms, namely CGA-A and CGA-B. Each item follows a specific grammatical category and sub-category. The mean item difficulty of CGA-A was 46.44 logits and that of CGA-B was 46.63 logits. Their overall item difficulties were very similar to each other. The difference is 0.19 logits. Among the 46 pairs of items under the same sub-category of the two alternate forms, 84.78% of them (39 pairs of items) had a difference of less than 4 logits. For the item pairs which had a larger difference, they were all positively endorsed by the 10 SMEs with ratings > 4.5 on their appropriateness and relevance, except one item under the category of "Binding" (with an item code "bnrfPS07") had a rating of 4.1. They were kept at this stage for a wider coverage of grammatical sub-categories. Further review would be made after different validity and reliability measures, especially the results of the Alternate Forms Reliability, which helps to determine if the two alternate forms were having high equivalence and correlation between each other.

Table 21. The items selected for the two alternate forms, namely CGA-A and CGA-B (item difficulty based on 831 TD data)

Item	Code	Grammatical Category	CGA-A (Item Code)	Item Difficulty (logit)	CGA-B (Item Code)	Item Difficulty (logit)
1	*S01*	*ba-construction*	babvGJ03	53.09	babvGJ04	55.82
2	*S01*	*ba-construction*	bacoGJ01	44.29	bacoGJ02	38.44
3	*S01*	*ba-construction*	baxxTV04	51.56	baxxTV03	52.07

4	*S02*	*passive*	beixTV04	44.76	beixTV02	46.11
5	*S02*	*passive*	bpspPM02	44.82	bpspPM03	41.74
6	*S03*	*binding*	bncrPS04	45.56	bncrPS01	43.87
7	*S03*	*binding*	bnpnPS01	47.18	bnpnPS03	51.00
8	*S03*	*binding*	bnpnPS07	42.28	bnpnPS08	42.87
9	*S03*	*binding*	bnrfPS04	41.32	bnrfPS02	40.82
10	*S03*	*binding*	bnrfPS05	36.67	bnrfPS07	45.22
11	*S04*	*relative clause*	rcooPS01	45.89	rcooPS02	46.06
12	*S04*	*relative clause*	rcosPS02	50.52	rcosPS03	50.47
13	*S04*	*relative clause*	rcsoPS01	50.90	rcsoPS04	48.52
14	*S04*	*relative clause*	rcssPS04	47.96	rcssPS03	47.81
15	*S05*	*comparatives*	cnbbPM01	44.76	cnbbPM02	45.67
16	*S05*	*comparatives*	cnbbPM05	47.86	cnbbPM06	48.37
17	*S05*	*comparatives*	cnmyPM03	44.41	cnmyPM04	44.11
18	*S05*	*comparatives*	compPS03	42.21	compPS04	41.25
19	*S06*	*quantification*	nqnqPS03	48.42	nqnqPS04	48.62
20	*S06*	*quantification*	nqqnPS02	42.80	nqqnPS03	41.25
21	*S06*	*quantification*	qualTV04	52.54	qualTV01	49.16
22	*S06*	*quantification*	quevTV04	47.39	quevTV02	44.99
23	*S07*	*double-object construction*	docxWR02	47.13	docxWR03	45.05
24	*S08*	*locative existential*	locaWR01	55.77	locaWR02	56.18
25	*S08*	*locative existential*	lociWR01	48.32	lociWR02	50.47
26	*S09*	*control*	ctocPS03	40.60	ctocPS04	42.08
27	*S10*	*cleft sentence*	clseSC02	49.07	clseSC04	49.65
28	*S11*	*question*	qmmaSC03	55.36	qmmaSC01	55.09
29	*S11*	*question*	qmreSC04	46.81	qmreSC01	44.88
30	*S12*	*morpheme distinction*	mdeiFB02	44.58	mdeiFB04	44.41
31	*S12*	*morpheme distinction*	mdexFB04	45.95	mdexFB03	46.87
32	*S12*	*morpheme distinction*	mdixFB04	42.35	mdixFB02	49.51
33	*S13*	*negation*	negbFB03	42.14	negbFB04	49.85
34	*S13*	*negation*	negmFB03	41.46	negmFB02	39.52
35	*S14*	*preposition*	precFB03	42.14	precFB04	42.80
36	*S14*	*preposition*	predFB01	56.64	predFB02	57.37
37	*S14*	*preposition*	pregFB02	51.04	pregFB01	51.98
38	*S14*	*preposition*	prexFB04	51.47	prexFB03	52.54
39	*S14*	*preposition*	prezFB04	49.12	prezFB03	54.50
40	*S15*	*localizer*	loloGJ01	47.81	loloGJ04	44.88
41	*S16*	*aspect*	aspfTV08	45.50	aspfTV07	38.69
42	*S16*	*aspect*	aspgTV04	45.33	aspgTV03	39.91
43	*S17*	*question words*	qwadFB04	45.95	qwadFB03	49.70
44	*S17*	*question words*	qwarFB02	45.28	qwarFB04	44.23
45	*S18*	*question particle*	qpmaGJ04	39.99	qpmaGJ03	39.36
46	*S18*	*question particle*	qpneGJ01	39.19	qpneGJ03	41.11
			Mean:	**46.44**	**Mean:**	**46.63**

The measures of the two alternate forms were also grouped under the 18 grammatical categories for another way of comparisons between the two lists of items. The mean item difficulty (in logits) in each grammatical category was calculated and summarized in Figure 8 and Table 22. As observed, among the 18 grammatical categories, "Control" is the easiest grammatical category and "Locative Existential" is the most difficult category based on the 831 TD primary school students.

Table 22. The mean item difficulty of CGA-A and CGA-B in 18 grammatical categories (N=831 TD data)

Grammatical Categories	**CGA-A (logits)**	**CGA-B (logits)**	**Mean (logits)**	**Difference (A-B) (logits)**
S01 Ba-construction	49.65	48.78	49.21	0.87
S02 Passive	40.72	45.67	43.19	-4.95
S03 Binding	44.23	44.06	44.15	0.17
S04 Relative clause	48.82	48.22	48.52	0.60
S05 Comparatives	44.81	44.85	44.83	-0.04
S06 Quantification	45.61	44.94	45.27	0.67
S07 Double-object construction	47.13	45.05	46.09	2.08
S08 Locative existential	52.05	53.33	52.69	-1.28
S09 Control	40.60	42.08	41.34	-1.48
S10 Cleft sentence	49.07	49.65	49.36	-0.58
S11 Question	50.30	53.52	51.91	-3.22
S12 Morpheme distinction	44.29	46.93	45.61	-2.64
S13 Negation	41.80	44.69	43.24	-2.89
S14 Preposition	49.95	49.26	49.61	0.69
S15 Localizer	47.81	44.88	46.35	2.93
S16 Aspect	45.42	39.30	42.36	6.12
S17 Question words	45.62	46.97	46.29	-1.35
S18 Question particle	51.09	49.99	50.54	1.10

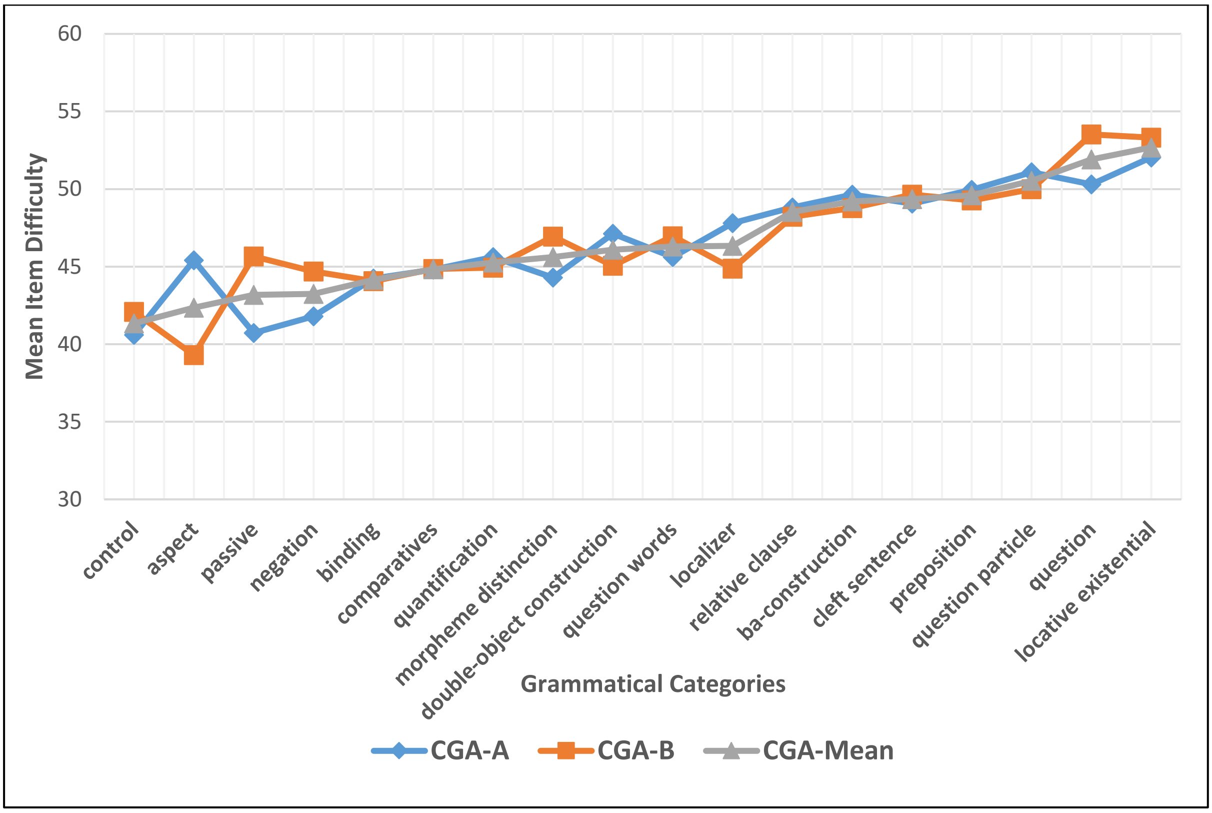


Figure 8. The mean item difficulty of the two alternate forms of CGA in 18 grammatical categories (N=831 TD data)

To ensure high equivalence between the two short forms, the mean item difficulty of the 46 different grammatical sub-categories were compared. Most of them showed to have comparable mean item difficulties. Mean item difficulties of Aspect, Passive and Questions had greater discrepancies between the two short forms. A follow-up investigation was made to check for SMEs' review on their relevance and appropriateness of the 6 items included in these 3 grammatical categories. All of them received a very positive endorsement from SMEs, with high average rating from 4.6-4.9. Thus, no deletion was done for these items.

6.1.2.2 Descriptive Statistics of the Two Alternate Forms

Table 23 summarized the results of the item analyses for the two CGA alternate forms based on either TD or DHH subjects. For TD subjects (n=831), the mean item difficulty of CGA-A and CGA-B are 46.44 logits (*SD*=4.46 logits) and 46.63 logits (*SD*=5.02 logits) based on the TD subjects. The mean difference between CGA-A and CGA-B was 0.19 logits. For DHH subjects (*N*=39), the mean item difficulty of CGA-A and CGA-B are 46.42 logits (*SD*=7.77logits) and 45.98 logits (*SD*=9.20 logits) based on the DHH subjects. The means and standard deviations were very similar between the two short forms. A larger standard deviation of the DHH data was observed when compared to that of the TD data, which means that DHH students in this study had greater individual differences in their CGA performance.

Table 23. Item Difficulty of CGA-A and CGA-B based on 831 TD and 39 DHH Subjects

	CGA-A (N=46 items)		CGA-B (N=46 items)	
	TD Subjects (*N*=831)	DHH Subjects (*N*=39)	TD Subjects (*N*=831)	DHH Subjects (*N=39*)
Range	36.67-56.64	30.25-65.25	38.44-57.37	15.63-64.40
Mean	46.44	46.42	46.63	45.98
SD	4.46	7.77	5.02	9.20

*Remark: The results were based on a joint analysis for the 92 items of the two equivalent lists

Paired sample t-test revealed that there was no significant difference between the mean item difficulty estimates of CGA-A and CGA-B based on both TD data (*t*= -0.411, *df*=45, *p*=0.683) and DHH data (*t*=0.549, *df*=45, *p*=.586). The results confirmed that CGA-A and CGA-B were having a comparable level of difficulty for both TD and DHH subjects.

As a summary, a student tested by the two equivalent short tests should give very similar, if not the same results. In another word, students' person ability estimates or their test scores based on the two alternate forms should be highly comparable and correlated with each other. The results of the above analyses confirmed that the two lists of CGA items had comparable levels of difficulties for both TD and DHH subjects. No significant difference between the could be found between the mean difficulty of CGA-A and CGA-B. In the following section, a series of reliability and validity measures would be conducted for the two forms of CGA. When the test was confirmed to be reliable and valid, the norms of the two CGA short tests could be set up for different practical reasons.

Chapter 7: Reliability and Validity of the Two Alternate Forms of CGA

7.1 Areas of Reliability and Validity Measures

To ensure that the two CGA short tests can accurately assess the targeted latent trait, that is the grammatical knowledge of written Chinese of the primary school students in Hong Kong, different reliability and validity measures were performed to further validate the two 46-item alternate forms. There were two stages of review performed. Besides using the database established by Tang, et al. (2023) from 2015-2019, validity and reliability measures were also conducted based on a new set of data collected from a regular primary school adopted the Sign Bilingualism and Co-enrollment in Deaf Education (SLCO) Programme in 2022 (the results will be reported in Chapter 8). In this programme, a relatively large group of DHH students were co-enrolled in a school with typically developing students. Because no repeated measures and other related assessments were conducted during the initial collection of the norming data, the test-retest reliability, and the convergent validity of the two alternate forms of CGA could only be tested using this new set of data.

In this chapter, we will report on the results of the reliability and validity measures based on the 831 TD and 39 DHH data collected from 2015-2019. After all the reliability and validity measures were completed, the 831 TD data were used to develop the norms for the two short tests and the 39 DHH from the same database were mainly used to collect more validity and reliability evidence of the two CGA short tests. The numbers of TD and DHH subjects and their grade levels used in this phase of study are summarized in the table below (see Table 24).

Table 24 Number of TD and DHH data at different grade levels used for the reliability and validity measures of the two alternate forms

Dataset	P1	P2	P3	P4	P5	P6	**Total**
TD norm data	84	170	216	100	159	102	**831**
DHH data	8	8	7	6	4	6	**39**

The results of reliability and validity measures for the two short versions of CGA, namely CGA-A and CGA-B, will be reported in the following chapters according to the following sequence:

1. Reliability Measures
   - i) Item/Person Separation Reliability
   - ii) Internal Consistency
   - iii) Alternate Forms Reliability
2. Validity Measures
   - i) Content Validity
   - ii) Known-Group Validity
   - iii) Construct Validity

7.2 Reliability Measures

In this section, the reliability of the two selected lists of 46-item CGA would be assessed with different measures. The item and person reliability as well as the separation statistics according to the Rasch analyses of both CGA-A and CGA-B will be reported. Then the two alternate forms' internal consistency reflected by the Cronbach's alpha were also reviewed based on the students' raw scores. At last, the Alternate Forms Reliability was conducted to review the equivalence of the two short forms of CGA.

7.2.1 Person/Item Separation Reliability

“Person separation indicates how efficiently a set of items is able to separate those persons measured. Item separation indicates how well a sample of people is able to separate those items used in the test” (Wright & Stone, 1999, p.151). The item and person separation reliability based on Rasch analyses were to see if the two established alternate forms were reliable for distinguishing persons with different abilities. To further review the equivalence of CGA-A and CGA-B, psychometric reviews on the short forms, i.e., CGA-A and CGA-B were conducted separately. Both the results from TD and DHH subjects will be reported.

7.2.1.1 Rasch Analysis of Norming Data from TD Subjects

Rasch analysis based on the 831 TD dataset was conducted for the two short forms separately (see Table 25 for a summary of the results). The results of person separation reliability were positive. The person reliability of CGA-A and CGA-B were both .86, in addition, the results of their person separation statistics were 2.44 and 2.49 respectively, which are both >2. Indeed, the results of reliability based on Rasch model also reflect high internal consistency of the items (Anselmi, Colledani, & Robusto, 2019).

The item reliability of CGA-A and CGA-B were both .98, which are very close to 1.0. According to Linacre (1995), the results indicated that the person sample was large enough to confirm the item difficulty hierarchy of the two forms. The results revealed that the items were reliable for repeated measures. In addition, their item separation values were 6.62 and 7.49 respectively, which are both >2. The results imply that the items of the two CGA short forms are having good reliability to distinguish students with different abilities. This will be further verified by investigating the significance of grade differences on CGA test results (see Section

7.3.2). The evidence of high item and person separation reliability implies reproducibility of their item and person measures in repeated test situations (Aryadoust, Ng, & Sayama, 2021). The results based on TD subjects provide important reliability evidence for the two short forms of CGA.

Table 25. Results of Rasch analysis separately for CGA-A and CGA-B (TD subjects)

	CGA-A		CGA-B	
	Person (*N*=831)	Item (*N*=46)	Person (*N*=831)	Item (*N*=46)
Range of logits	33.05-84.79	36.69-56.83	31.42-85.16	38.25-57.45
Mean	55.49	46.53	55.40	46.53
SD	8.69	4.44	8.92	5.04
Separation	2.44	6.62	2.49	7.49
Reliability	.86	.98	.86	.98

7.2.1.2 Rasch Analysis of Data from DHH Subjects

Though DHH data were not included in the establishment of the norms of the two finalized short tests, Rasch analysis was conducted based on the dataset with 39 DHH subjects to collect evidence for the validation of the two alternate forms. When the two alternate forms were reviewed on their person reliability, the results for both forms, as showed in Table 26, were .86. The values of person reliability >.80 indicated that the two alternate forms of CGA were reliable to provide consistent test results for the same group of persons (Linacre, 1995). The values of person separation for CGA-A and CGA-B were 2.51 and 2.43 respectively, which were both >2. The results indicate that the two alternate forms were able to distinguish DHH students with different levels of ability regarding their grammatical knowledge in written Chinese.

Item reliability of the CGA-A and CGA-B were found to be .80 and .81 respectively (both were < .80). Moreover, the values of item separation were 2.02 and 2.09 (both were < 2) (see Table 26). The results indicated a good item separation reliability of the two forms though there were only a small number of subjects included in the analysis (Linacre, 1995).

Table 26. Results of Rasch analysis separately for CGA-A and CGA-B (DHH subjects)

	CGA-A		CGA-B	
	Person (*N*=39)	Item (*N*=46)	Person (*N*=39)	Item (*N*=46)
Range of logits	38.21-87.57	29.94-66.16	41.86-87.61	15.65-64.18
Mean	56.91	46.53	56.44	45.86
SD	10.79	7.88	10.22	9.04
Separation	2.51	2.02	2.43	2.09
Reliability	.86	.80	.86	.81

As a summary of the results for item and person separation reliability of CGA-A and CGA-B, both forms of CGA are confirmed to be reliable in assessing Chinese grammatical knowledge of primary school students, with or without hearing loss.

7.2.2 Internal Consistency

To assess the internal consistency of the two 46-item alternate forms of CGA, Cronbach' alpha was conducted separately for the two forms using students' raw scores. The results based on TD subjects (*N*=831) were α = .90 for CGA-A (46 items) and α = .91 for CGA-B (46 items). The results were both ≥ .90, reflecting that both the two forms had excellent internal consistency. Excellent results regarding internal consistency were also got from the DHH data

(*N*=39). Their Cronbach's alphas were both ≥ .90, i.e., α = .92 for CGA-A (46 items) and α = .90 for CGA-B (46 items). In sum, the items of both forms of CGA are converging to the same direction to measure the targeted latent trait, which is the grammatical knowledge of students in written Chinese.

7.2.3 Alternate Forms Reliability

In this study, in order to develop two CGA short tests for future educational and clinical use, the test scores obtained from the two alternate forms were expected to be highly correlated and comparable with each other. Students' performance in the two short tests should ideally be the same, which means that the test scores of the same group of students tested by CGA-A and CGA-B should have no significant difference. In order to confirm the equivalence of the two forms, Rasch analysis was conducted for both forms and the measures of person ability were used to conduct analysis for Intra-class Correlation Coefficient (ICC) (Shrout & Fleiss, 1979; McGraw & Wong, 1996). In addition to the logit measures from Rasch analysis, the raw scores of students were also used for the review of their Alternate Forms Reliability.

To assess the Alternate Forms Reliability between the two alternate forms of CGA, ICCs were calculated using SPSS statistical package version 27 based on a single-measure, absolute-agreement, 2-way mixed-effects model (with reference to Koo & Li, 2016). The results of ICCs reflected not only the degree of correlation but also the agreement between the two measurements (Koo & Li, 2016).

7.2.3.1 Descriptive Statistics

Before reviewing the results of alternative forms reliability for CGA-A and CGA-B, the results of the students assessed by the two lists were compared in logits estimated by Rasch analyses. Table 27 summarizes the results of TD and DHH students assessed by CGA-A and CGA-B. As the two short versions of CGA are developed for the use of educational and rehabilitation professionals and the norms would be developed based on students' raw scores, with reference to Holmefur, Aarts, Hoare, and Krumlinde-Sundholm (2009), the analyses were also conducted based on the students' raw scores.

Table 27 shows that the means and standard deviations of the two lists were quite similar no matter in terms of logits (CGA-A: *M*=55.49 logits, *SD*=8.69; and CGA-B: *M*=55.40 logits, *SD*=8.92) or raw scores (CGA-A: *M*=32.79, *SD*=8.76; and CGA-B: *M*=32.51, *SD*=8.85), moreover, the range of measures projected by CGA-A and CGA-B were also very similar in either logits or raw scores.

For DHH subjects, the mean person ability estimated from CGA-A and CGA-B were 56.91 logits (*SD*=10.79) and 56.44 logits (*SD*=10.22) respectively. Their results were also similar in terms of raw scores (CGA-A: *M*=32.69, *SD*=9.20; and CGA-B: *M*=32.74, *SD*=8.42) (see Table 27). In both conditions, the means and standard deviations of CGA-A and CGA-B were similar, no matter in logits or raw scores, only that CGA-A had a slightly wider range of scores than that of CGA-B.

Comparing the results between the TD and the DHH subjects (see Table 27), the mean scores of DHH subjects were a little bit higher than that of the TD subjects, but the standard deviations of the results of DHH subjects were also greater than that of the TD subjects. According to the results, it seems that the DHH subjects in this study had greater individual differences than the TD subjects.

Table 27. Descriptive statistics for the results of TD and DHH subjects by the two alternate forms of CGA in both logits and raw scores

Person Ability	TD Subjects (*N*=831)		DHH Subjects (*N*=39)	
	CGA-A	CGA-B	CGA-A	CGA-B
Logits				
Range	33.05-84.79	31.42-85.16	38.21-87.57	41.86-87.61
Mean (SD)	55.49 (8.69)	55.40 (8.92)	56.91 (10.79)	56.44 (10.22)
Raw Scores				
Range	7-46	6-46	13-46	18-46
Mean (SD)	32.79 (8.76)	32.51 (8.85)	32.69 (9.20)	32.74 (8.42)

7.2.3.2 Intra-class Correlation Coefficients (ICC)

To further confirmed the Alternate Forms Reliability of the two short forms of CGA, statistical analyses were conducted in terms of intra-class correlation coefficients (ICC). ICC has a value range from 0 to 1. A higher value of ICC indicates a higher degree of agreement. As suggested by Koo and Li, (2016), a coefficient less than .50 represents poor reliability. An ICC values between .50-.75 is moderate, values between .75-.90 is good and values greater than .90 represents excellent reliability. The levels of reliability were reviewed based on the above-mentioned criteria as well as the lower bound and upper bound of the 95% confidence intervals of all ICC estimates (Koo & Li, 2016). In this study, the lowest acceptable ICC value was .80 for all results, expecting a “good” or “excellent” alternate forms reliability between the two short tests.

7.2.3.3 Alternate Forms Reliability Based on TD Subjects

In the paragraphs below, we will discuss the results of the data from typically developing students first. Then the discussion will focus on DHH subjects. Table 28 and

Table 29 show the results of ICC measures for the two CGA lists based on both logit measures and raw scores of typically developing (TD) subjects. The reliability measures were conducted for the two short forms based on the whole group of TD subjects (*N*=831) first and then separately for different grade levels. Moreover, Table 30 and Table 31 summarizes the corresponding results of ICC measures based on the data from deaf and hard-of-hearing (DHH) subjects. A discussion of the results will be made after reporting the results for both groups of subjects.

Results in Table 28 and Table 29 indicates a “good” to “excellent” intra-class reliability between the two forms based on the assessment results of the 831 TD subjects, no matter in logit measures or raw scores (according to the criteria proposed by Koo & Li, 2016, p. 161). The values of reliability were .886 (with the 95% confidence intervals between .870-.900) based on logit measures and the reliability coefficient was .918 (with the 95% confidence intervals between .907-.928) based on raw score totals. The ICC values based on raw score totals were .90 in a 95% confidence interval between .890-.910. F-test results based on the 831 TD data for all grade levels were significant for both logit measures (*F*=16.498, *df*1=830, *df*2=830, *p*<.001) and raw score totals (*F*=23.508, *df*1=830, *df*2=830, *p*<.001).

When the ICCs were reviewed by grade levels, the results ranged between .809-.870 based on logits, and .865-.928 based on raw scores. All the reliability coefficients were greater than .80, reflecting a good reliability between the two lists for all grade levels. Significant results were got from all the *F*-test with *p*-values<.001 for all six grade levels.

Table 28. Results of intra-class correlation coefficients (ICC) for alternate forms reliability of CGA-A and CGA-B based on logits (TD Subjects)

Alternate Forms Reliability#	Intraclass	95% Confidence Interval		F Test with True Value 0			
(Single Measures : Logits)	Correlation	Lower Bound	Upper Bound	Value	*df* 1	*df* 2	Sig.
CGA-A vs CGA-B (P1-P6; n=831)	.886	.870	.900	16.50	830	830	.000
CGA-A vs CGA-B (P1; n=84)	.869	.806	.913	14.18	83	83	.000
CGA-A vs CGA-B (P2; n=170)	.870	.829	.903	14.39	169	169	.000
CGA-A vs CGA-B (P3; n=216)	.841	.797	.876	11.53	215	215	.000
CGA-A vs CGA-B (P4; n=100)	.809	.728	.867	9.37	99	99	.000
CGA-A vs CGA-B (P5; n=216)	.836	.783	.878	11.17	158	158	.000
CGA-A vs CGA-B (P6; n=102)	.843	.775	.891	11.60	101	101	.000

#ICC estimates were calculated based on a single-measures, absolute-agreement, 2-way mixed-effects model

Table 29. Results of intra-class correlation coefficients (ICC) for alternate forms reliability of CGA-A and CGA-B based on raw scores (TD subjects)

Alternate Forms Reliability#	Intraclass	95% Confidence Interval		F Test with True Value 0			
(Single Measures : Raw Scores)	Correlation	Lower Bound	Upper Bound	Value	*df* 1	*df* 2	Sig.
CGA-A vs CGA-B (P1-P6; n=831)	.918	.907	.928	23.508	830	830	.000
CGA-A vs CGA-B (P1; n=84)	.885	.829	.924	16.300	83	83	.000
CGA-A vs CGA-B (P2; n=170)	.875	.834	.906	14.957	169	169	.000
CGA-A vs CGA-B (P3; n=216)	.891	.859	.915	17.323	215	215	.000
CGA-A vs CGA-B (P4; n=100)	.865	.805	.907	13.660	99	99	.000
CGA-A vs CGA-B (P5; n=216)	.890	.852	.918	17.293	158	158	.000
CGA-A vs CGA-B (P6; n=102)	.928	.896	.951	27.228	101	101	.000

#ICC estimates were calculated based on a single-measures, absolute-agreement, 2-way mixed-effects model

7.2.3.4 Results Based on DHH Subjects

As showed in Table 30 and Table 31, the alternate forms reliability of the two alternate forms was also considered “good” to “excellent” based on the analysis of intra-class correlation coefficients using DHH subject data. The ICCs based on all 39 DHH subjects were .941 (with the 95% confidence intervals between .890-.968) based on the logit measures and .936 (with the 95% confidence intervals between .881-.966) based on the raw scores. Analysis of variance showed that the ICC measures were all significant no matter the results were in logit measures (*F*=32.791, *df*1=38, *df*2=38, *p*<.001) or raw score (*F*=29.440, *df*1=38, *df*2=38, *p*<.001) (see both Table 30 and Table 31).

Table 30. Results of intra-class correlation coefficients (ICC) for alternate forms reliability of CGA-A and CGA-B based on logits (DHH Subjects)

Alternate Forms Reliability#	Intraclass	95% Confidence Interval		F Test with True Value 0			
(Single Measures : Logits)	Correlation	Lower Bound	Upper Bound	Value	*df* 1	*df* 2	Sig.
CGA-A vs CGA-B (P1-P6; n=39	.941	.890	.968	32.791	38	38	.000
CGA-A vs CGA-B (P1; n=8)	.779	.240	.951	7.479	7	7	.008
CGA-A vs CGA-B (P2; n=8)	.889	.556	.977	15.595	7	7	.001
CGA-A vs CGA-B (P3; n=7)	.916	.562	.985	30.512	6	6	.000
CGA-A vs CGA-B (P4; n=6)	.941	.683	.991	30.888	5	5	.001
CGA-A vs CGA-B (P5; n=4)	.971	.618	.998	51.660	3	3	.004
CGA-A vs CGA-B (P6; n=6)	.914	.535	.987	19.708	5	5	.003

#ICC estimates were calculated based on a single-measures, absolute-agreement, 2-way mixed-effects model

Table 31. Results of intra-class correlation coefficients (ICC) for alternate forms reliability of CGA-A and CGA-B based on raw scores (DHH Subjects)

Alternate Forms Reliability#	Intraclass	95% Confidence Interval		F Test with True Value 0			
(Single Measures : Raw Scores)	Correlation	Lower Bound	Upper Bound	Value	*df* 1	*df* 2	Sig.
CGA-A vs CGA-B (P1-P6; n=39)	.936	.881	.966	29.440	38	38	.000
CGA-A vs CGA-B (P1; n=8)	.785	.303	.952	8.681	7	7	.005
CGA-A vs CGA-B (P2; n=8)	.898	.605	.978	18.087	7	7	.001
CGA-A vs CGA-B (P3; n=7)	.932	.670	.988	33.656	6	6	.000
CGA-A vs CGA-B (P4; n=6)	.916	.520	.988	19.566	5	5	.003
CGA-A vs CGA-B (P5; n=4)	.964	.661	.998	67.500	3	3	.003
CGA-A vs CGA-B (P6; n=6)	.868	.315	.980	12.209	5	5	.008

#ICC estimates were calculated based on a single-measures, absolute-agreement, 2-way mixed-effects model

The results based on the analysis of variance were all significant with *p*<0.01, no matter the results were estimated or calculated in logits or raw scores (see Table 30 and Table 31), but the intra-class coefficient coefficients (ICC) reviewed by individual grade levels varied greatly, possibly because of the small sample size of DHH students at each grade levels (from *N*=4 to *N*=8). The ICC results ranged between .779-.971 based on logit measures and 0.785-0.964 based on raw scores. The lowest ICC value between CGA-A and CGA-B was from the data of P1 DHH subjects, which was .779 (with a 95% confidence interval between .240-.951) based on logit measures and .789 (with a 95% confidence interval between .303-.952) based on raw scores.

The range of intraclass correlation coefficients was wide. The results could not be interpreted simply from the value of the coefficient. According to Koo and Li (2016), the alternate forms reliability was considered "good" (ICC=.779 based on logits and .785 based on raw scores) according to the reliability coefficient. However, when the 95% confidence interval was considered, the true ICC value might land on any point between .240-.951 based on the results in logits or .303-.952 based on the results in raw scores. The alternate forms reliability between CGA-A and CGA-B based on the results of P1 DHH subjects could be considered "poor" in one extreme and "excellent" to another. The results based on the should be interpreted with reservation. As the norms were built upon the results of TD subjects. The results for DHH subjects could be taken as a reference in this study.

The alternate forms reliability between the two lists of CGA items was assessed based on different variables, including subject groups (TD versus DHH subjects), types of scores (test results in "logit measures" versus "raw scores"), and grade levels (subjects from "all grade levels as a whole" versus "individual grade levels from P1-P6"). As a whole, the items in CGA-

A and CGA-B are highly correlated with each other. The values of ICCs as well as the 95% confidence intervals basically fall into the range of “.75-.90” or “ >.90”, confirming that both CGA-A and CGA-B possess a “good-to-excellent” alternate forms reliability. Though the results of DHH data by different grade levels have a wider range of results when compared to that of the TD data, small sample size of DHH subjects may be a reason that affects the statistical findings. In sum, the results of alternate forms reliability reflect that the two short forms of CGA are comparable and equivalent to each other. They are reliable to be used interchangeably for the assessment of students’ grammatical knowledge in written Chinese with when the norms have been set up.

7.2.4 Interim Discussion

Reliability is defined as the extent to which measurements can be replicated. In this study, we assessed the reliability of the two short versions of CGA through different statistical analyses. By investigating the item and person reliability and their separation statistics through Rasch analysis, the items of the two alternate forms are found to be reliable in discriminating primary school students with different levels of abilities. The good person separation reliability of the two lists also reflect that the results of CGA-A and CGA-B are likely to be replicable and consistent though further review is required.

Intraclass correlation coefficient (ICC) is a widely used reliability index in test-retest, intrarater, and interrater reliability analyses (Koo & Li, 2016). In this part, ICC is used to review the reliability between the two alternate forms, and the results show that the two lists are having “good to excellent” reliability, and able to give equivalent results for the same subject.

7.3 Validity Measures

After reporting some results of the reliability measures for the two alternate forms of CGA, we will report the results of different validity measures conducted for them. The analyses were based on the norming data with 831 TD subjects.

Different areas of validity measures were conducted including a review of the content validity of the two selected lists of items for the alternate forms of CGA. To provide evidence for validity of the two forms, the analysis for known-group validity was conducted with an assumption that students at a higher grade level would have better grammatical knowledge in written Chinese. In another words, the grade level is expected to be a significant factor affecting the students' performance in CGA-A and CGA-B. This part of review is also considered a review on the discriminative validity of the two alternate forms, determining if they are able to discriminate students with different abilities.

7.3.1 Content Validity of the Two Equivalent Lists

Based on the 18 grammatical categories, 46 sub-categories were selected for item development, and eventually 172 test items were generated for the initial version of CGA for the collection of data for validation and norm setting. As mentioned in Chapter 4, 10 Subjects Matter Experts (SMEs) were involved in an expert panel to review the content of the 172-item CGA. The SMEs filled in the questionnaire on the platform for the review panel (see Appendix C) regarding the administration and operations as well as different content areas of CGA. Besides the ratings they gave for the different aspects of review, they also provided written comments for the further development of CGA.

Regarding the review of the representativeness, relevance and appropriateness of CGA, there was a report in Chapter 4. As the representativeness of the 18 grammatical categories was fully endorsed by the SMEs (Mean CVI=.90), the items selected for the two alternate forms were all from these categories.

In this section, the content validity of the two 46-item alternate forms of CGA would be reviewed separately based on the results content validation by the review panel. In this part, the review aims to ensure that the 92 selected items for the two 46-item alternate forms have good content validity. The results based on ratings given by the SMEs and the projected Content Validity Index (CVI) were thus summarized and reviewed accordingly (see Table 32). Analysis of Intra-class correlation coefficient (ICC) was also conducted to see if the SMEs' ratings for the items of the two lists were comparable and correlated well with each other.

Table 32. Summary of the Mean Ratings and CVIs on the Two 46-item Equivalent Lists

	Ratings on Appropriateness		Ratings on Relevance	
Equivalent Lists	CGA-A (*N*=46)	CGA-B (*N*=46)	CGA-A (*N*=46)	CGA-B (*N*=46)
Mean rating (SD)	4.67 (0.19)	4.63 (0.20)	4.75 (0.15)	4.74 (0.16)
Mean CVIs (SD)	.92 (.08)	.91 (.09)	.96 (.06)	.95 (.07)

The mean ratings (CVIs) of the 10 SMEs regarding the appropriateness of items were 4.67 (CVI=.92) and 4.63 (CVI=.91) for CGA-A and CGA-B respectively (see Table 32). The results in terms of the CVIs of CGA-A and CGA-B were all > .90, representing that the items selected were appropriate for the assessment, with reference to the recommendations from Cannon &

Hubley (2014) on content validity measures.

Regarding the relevance of the items for the assessment, the mean ratings and CVIs of the SMEs were 4.75 (CVI=.96) and 4.74 (CVI=.95) for CGA-A and CGA-B respectively. Both the mean ratings for the two alternate forms were >4.5 and their CVIs were .92 and .90 respectively, representing that the two sets of items were highly endorsed by the SMEs as having very good content validity.

Table 33. Intraclass Correlation Coefficients (ICC) between the content validity ratings for CGA-A and CGA-B on the appropriateness and relevance of items

Content Validation for items	Intraclass Correlation	95% Confidence Interval		F Test with True Value 0			
		Lower Bound	Upper Bound	Value	df1	df2	Sig.
Appropriateness: CGA-A vs. CGA-B	.697	.512	.827	5.861	45	45	.000
Relevance: CGA-A vs. CGA-B	.827	.708	.900	10.523	45	45	.000

Based on a single-measure, absolute-agreement, two-way mixed-effects model, the Intraclass Correlation coefficients (ICCs) summarized in Table 33 reflects that the intrarater reliability between the two lists was “moderate” (between .50-.75) for the SMEs’ ratings on the appropriateness of the items and “moderate to good” (between .50-.90) for the ratings on the items’ relevance.

In sum, the appropriateness and relevance of the items selected for CGA-and CGA-B from the 172-item pool were both positively endorsed by the SMEs according to the high mean ratings (>4.50) and CVIs (>.90). To further confirm the equivalence of the two alternate forms, we also looked into the intrarater reliability of the 10 SMEs on the items of the two short forms, and the results are positive regarding the items’ relevance and appropriateness. The above-mentioned evidence supports the claim that both lists of CGA items are having high content validity with moderate to good intrarater reliability between the two lists.

7.3.2 Known-Group Validity

Known-group validity is a measure that contributes to the construct validity of a measurement. It helps verify if an assessment tool is measuring what it intends to (Mooi & Sarstedt, 2011). As CGA is developed for assessing primary school students' grammatical knowledge in written Chinese. Students in school receiving Chinese language education should have continuous development in their Chinese grammatical knowledge. Therefore, the assessment should be reliable and robust enough to distinguish students with different levels of grammatical knowledge and their level of abilities should be associated with their grade levels. Measures for known-group validity in this study aims to see if the students at higher-grade levels will get a better result (no matter in logit measures or raw scores) in both short versions of CGA than those at lower grade levels.

7.3.2.1 Known-group Validity Based on TD Data

In order to test for the known-group validity of the two lists, one-way analysis of variance (ANOVA) was used to check if both CGA-A and CGA-B were able to distinguish among students with different grade levels. According to the mean plots showed in Figure 9 and Figure 10, the test scores of CGA-A and CGA-B were showing an increasing trend that was associated with the students' grade level. Further statistical investigations were then conducted to further confirm the results (see Table 34 and Table 35).

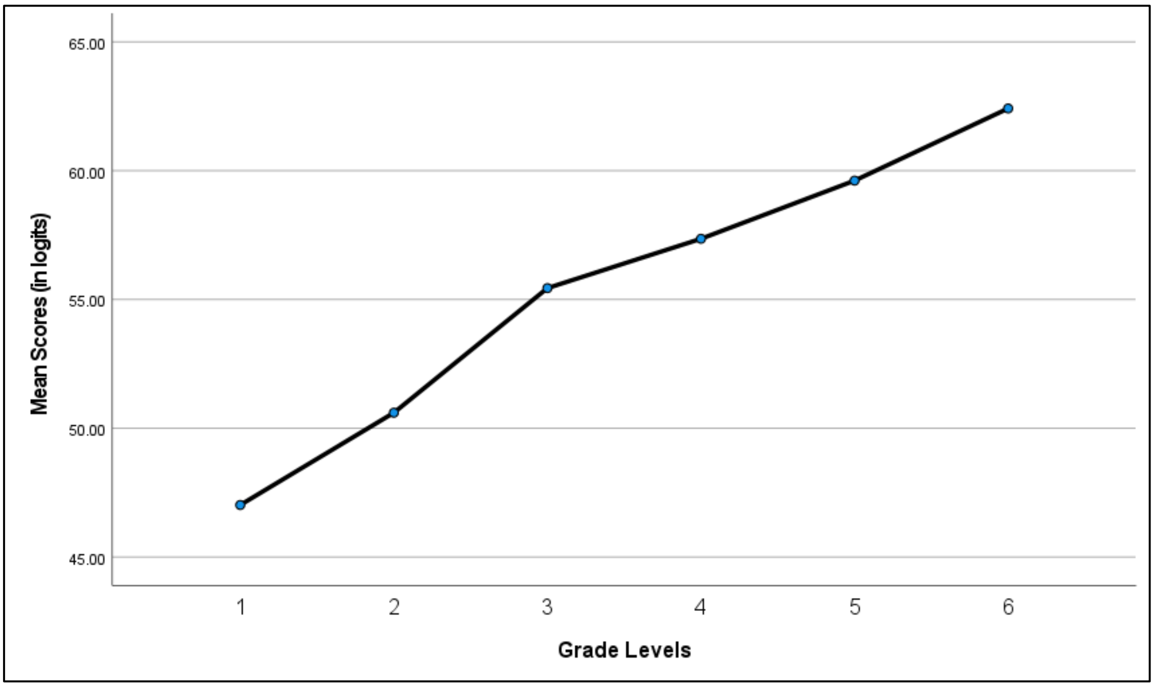


Figure 9. Mean plots for CGA-A by grade levels (TD Subjects)

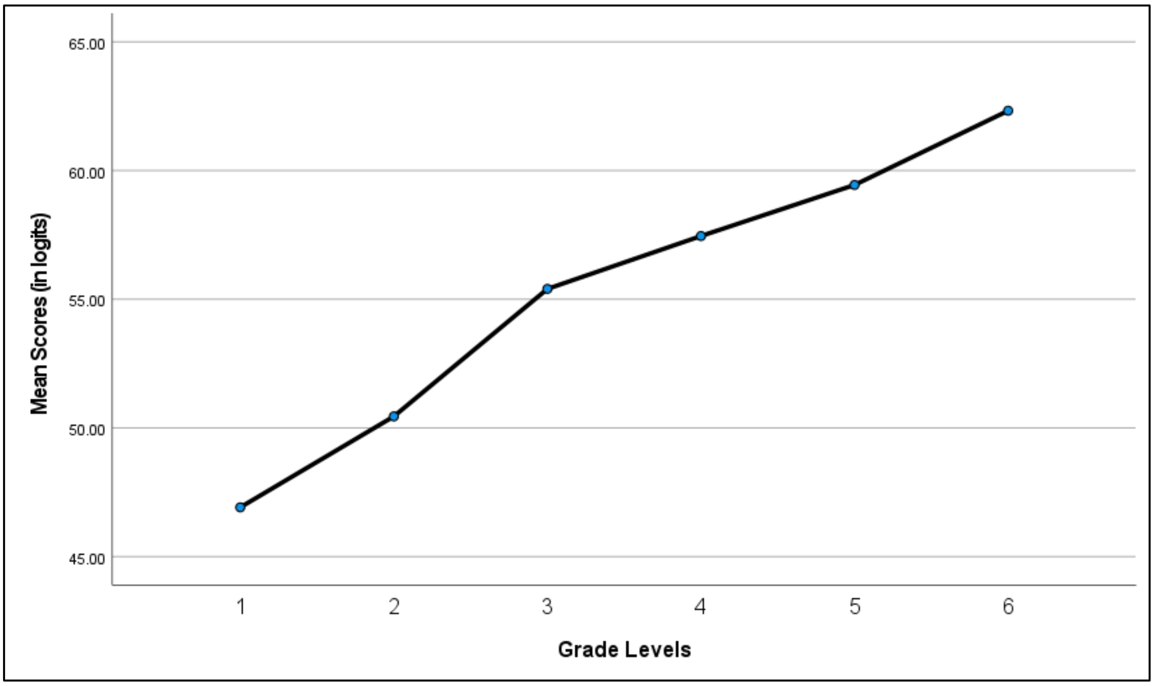


Figure 10. Mean plots for CGA-B by grade levels (TD Subjects)

The Shapiro-Wilk test of normality was conducted based on the data (in logits) and the results indicated that the assumption of normal distribution was violated with a significant level of *p*<.001 for all two sets of performance data, one dataset from CGA-A and the other from CGA-B. In addition, the two datasets were reviewed by the Levene's test, and the homogeneity of variance assumption was also found violated with a significant level of *p*<.05 for both CGA-A (*p*=.013) and CGA-B (*p*=.015). In this regard, the Welch's *F*-test was used for the analysis of variance and the Games-Howell test was used for the post hoc procedures (Mooi, & Sarstedt, 2011). The pairwise comparisons were conducted to determine if there were significant mean differences between the CGA scores of students at different grade levels. An alpha level of .05 was used for the subsequent analyses.

Results of one-way ANOVA using the Welch's *F*-Test confirmed that there was a significant main effect of grade level on CGA scores in logits with the results of Welch's *F*(5, 324.43)=69.96, *p*<.001, ω^2=.28 for CGA-A, and Welch's *F*(5, 325.55)=66.60, *p*<.001, ω^2=.27 for CGA-B. To further investigate the grade differences between CGA scores in more details, post hoc Games-Howell tests were conducted for pairwise comparisons among the students at different grade levels (see Table 34 and Table 35 for the results of CGA-A for CGA-B respectively). Results showed that the mean difference for CGA-A and CGA-B were statistically significant between P1-P2 (mean difference of CGA-A, *M*=-3.58, *p*<.001; and CGA-B, *M*=-3.53, *p*<.001) and P2-P3 (mean difference of CGA-A, *M*=-4.84, *p*<.001; and CGA-B, *M*=-4.95, *p*<.001) in both lists. Though there were no significant difference found between P3-P4, P4-P5 and P5-P6 in both CGA-A and CGA-B, the mean difference was found significant between P3 and P5 (mean difference of CGA-A, *M*=-4.17, *p*<.001; and CGA-B, *M*=-3.70, *p*<.001) as well as P4 and P6 (mean difference of CGA-A, *M*=-5.06, *p*<.001 and CGA-B, *M*=-3.40, *p*<.001). The results indicated that students required two years' time to show significant improvement in their CGA performance after P3.

Table 34. Results of Post-hoc Games-Howell Test for students tested by CGA-A at different grade levels (TD subjects, *N*=831)

			Mean Difference (Logits)				
Grade	N	Mean (SD)	P1	P2	P3	P4	P5
P1	84	47.02 (6.64)					
P2	170	50.60 (5.99)	-3.58**				
P3	216	55.44 (7.45)	-8.42**	-4.84**			
P4	100	57.36 (8.09)	-10.33**	-6.75**	**-1.92**		
P5	159	59.62 (7.61)	-12.59**	-9.01**	-4.17**	**-2.26**	
P6	102	62.42 (8.60)	-15.39**	-11.81**	-5.97**	-5.06**	**-2.80**

* *p*<.05, ** *p*<.01

Table 35. Results of Post-hoc Games-Howell Test for students tested by CGA-B at different grade levels (TD subjects, *N*=831)

			Mean Difference (Logits)				
Grade	N	Mean (SD)	P1	P2	P3	P4	P5
P1	84	46.92 (6.71)					
P2	170	50.45 (6.09)	-3.53**				
P3	216	55.40 (7.94)	-8.49**	-4.95**			
P4	100	57.46 (7.88)	-10.54**	-7.01**	**-1.98**		
P5	159	59.44 (7.81)	-12.53**	-8.99**	-3.70**	**-1.71**	
P6	102	62.33 (9.23)	-15.41**	-11.88**	-5.39**	-3.40**	**-1.69**

**p*<.05, ***p*<.01

According to the results mentioned above, grade level has a main effect on CGA performance. There showed a significant one-year difference between junior primary levels (from P1 to P3).

However, from P4 onward, one-year difference was not significant, but a two-year difference was significant.

7.3.2.2 Known-group Validity Based on DHH Data

Though the norms for the two CGA short tests would only be based on the data from typically developing (TD) students, the effect of grade level on CGA scores were also tested for data from DHH subjects. The sample size was small, but a significant main effect was still expected. Similar to the testing procedures for data of TD subjects, the one-way ANOVA was used to test for the hypothesis.

The Shapiro-Wilks test for normality indicated the data were statistically normal. The Levene's statistics revealed that the homogeneity of variance assumption was met, *F*(3,35)=0.19, *p*=.905 for CGA-A and *F*(3,35)=1.51, *p*=.228. One-way ANOVA was conducted for DHH student's grade level on their CGA scores (in logits) was conducted. The main effect of grade levels on CGA scores was significant for both CGA-A with *F*(5, 33)=5.51, *p*<.01 with an effect size of 0.46 and CGA-B with *F*(5, 33)=6.97, *p*<.01 with an effect size of 0.51. Grade level is a significant factor affecting CGA performance of DHH students, but the mean differences between adjacent grade levels could not be clearly identified (see Table 36 and Table 37). By observing the performance of the DHH students at different grade levels, there is a general trend that the mean logits increased with grade levels, from *M*=45.89 (CGA-A) and *M*=46.73 (CGA-B) for P1, growing up to *M*=64.25 (CGA-A) and *M*=63.09 (CGA-B) for P6. The mean differences between adjacent grade levels were generally small. The effect was not significant in most of the pairwise comparisons between grade levels based on the Tukey test. Significant mean differences were found for grade levels between P1-P4 or P6 and P2-P4 or P6, with the values ranged from -16.37 to -18.36, *p*<.01.

Table 36. Results of Post-hoc Tukey Test for students tested by CGA-A at different grade levels in Logits (DHH subjects, *N*=39)

			Mean Difference (Logits)				
Grade	N	Mean (SD)	P1	P2	P3	P4	P5
P1	8	45.89 (6.90)					
P2	8	51.15 (6.22)	**-5.27**				
P3	7	55.57 (7.42)	**-9.68**	**-4.42**			
P4	6	63.90 (5.00)	-18.02**	-12.75**	**-8.33**		
P5	4	58.77 (6.68)	**-12.88**	**-7.61**	**-3.20**	**-5.18**	
P6	6	64.25 (14.01)	-18.36**	-13.10**	**-8.68**	**-0.35**	**-5.49**

* *p*<.05, ** *p*<.01

Table 37. Results of Post-hoc Tukey Test for students tested by CGA-B at different grade levels in Logits (DHH subjects, *N*=39)

			Mean Difference (Logits)				
Grade	N	Mean (SD)	P1	P2	P3	P4	P5
P1	8	46.73 (4.53)					
P2	8	50.62 (4.82)	**-3.90**				
P3	7	53.44 (9.43)	**-6.71**	**-2.81**			
P4	6	63.36 (4.53)	-16.63**	-12.74**	**-9.92**		
P5	4	58.91 (5.76)	**-12.18**	**-8.28**	**-5.47**	**4.45**	
P6	6	63.09 (9.64)	-16.37**	-12.47**	**-9.65**	**0.26**	**-4.18**

* *p*<.05, ** *p*<.01

The individual difference between this group of DHH students was prominent. As observed from the data, the P4 DHH students (*M*=63.90 for CGA-A and *M*=63.36 for CGA-B) in this group of DHH students outperformed those at P5 or even P6. There may be factors like students' degree of hearing loss, speech perception ability, use of hearing device, early oral language development or age of early intervention contributing for the individual differences.

In sum, the two short tests, no matter CGA-A or CGA-B, showed significant grade differences on students' CGA performance. The two short versions of CGA were more effective in discriminating students' level of Chinese grammatical knowledge among the three junior primary grade levels than that of the students at the senior grade levels (from P3 onward). The mean differences between the CGA scores of students from P3-P6 were not significant for pairwise comparisons between adjacent grade levels, but a significant result could be found for two-year differences like P3-P5 or P4-P6. With the existing test items, the two CGA has a better discriminating power for students' CGA performance at junior primary than that at their senior primary levels.

7.3.3 Construct Validity

Construct-related validity mainly concerns whether the assessment is precisely testing the targeted latent trait, which is the grammatical knowledge in written Chinese. To achieve this, how the assessment and the items are developed is crucial. The involvement of different subject experts to help review the assessment based on a concrete theoretical construct can guarantee a better construct validity (Ng, 2014). During the development of CGA, the items were first developed based on comprehensive literature review. When all the items were ready, they were critically reviewed by three renowned scholars and researchers in Chinese linguistics and language acquisition. They provided expert opinions for item revision or refinement. In this study, the content validity of the assessment was also reviewed by five professional speech therapists and five teachers who had experience in Chinese language teaching for both TD and DHH students. The positive results in content validity (see Section 7.3.1) also support construct validity of CGA.

7.3.3.1 Evidence from Known-group Validity Measure

Results of known-group validity for the two CGA alternate forms (see Section 7.3.2) provided good evidence to confirm construct validity of the assessment. As CGA is developed to see how well primary school students comprehend grammatical structures in written Chinese, when CGA scores from the two tests are able to distinguish students at different grade levels (students at higher grade levels have better CGA scores), it reflects that the assessment has a good prediction of its intended results. Further validity testing results will also be reported in later sections.

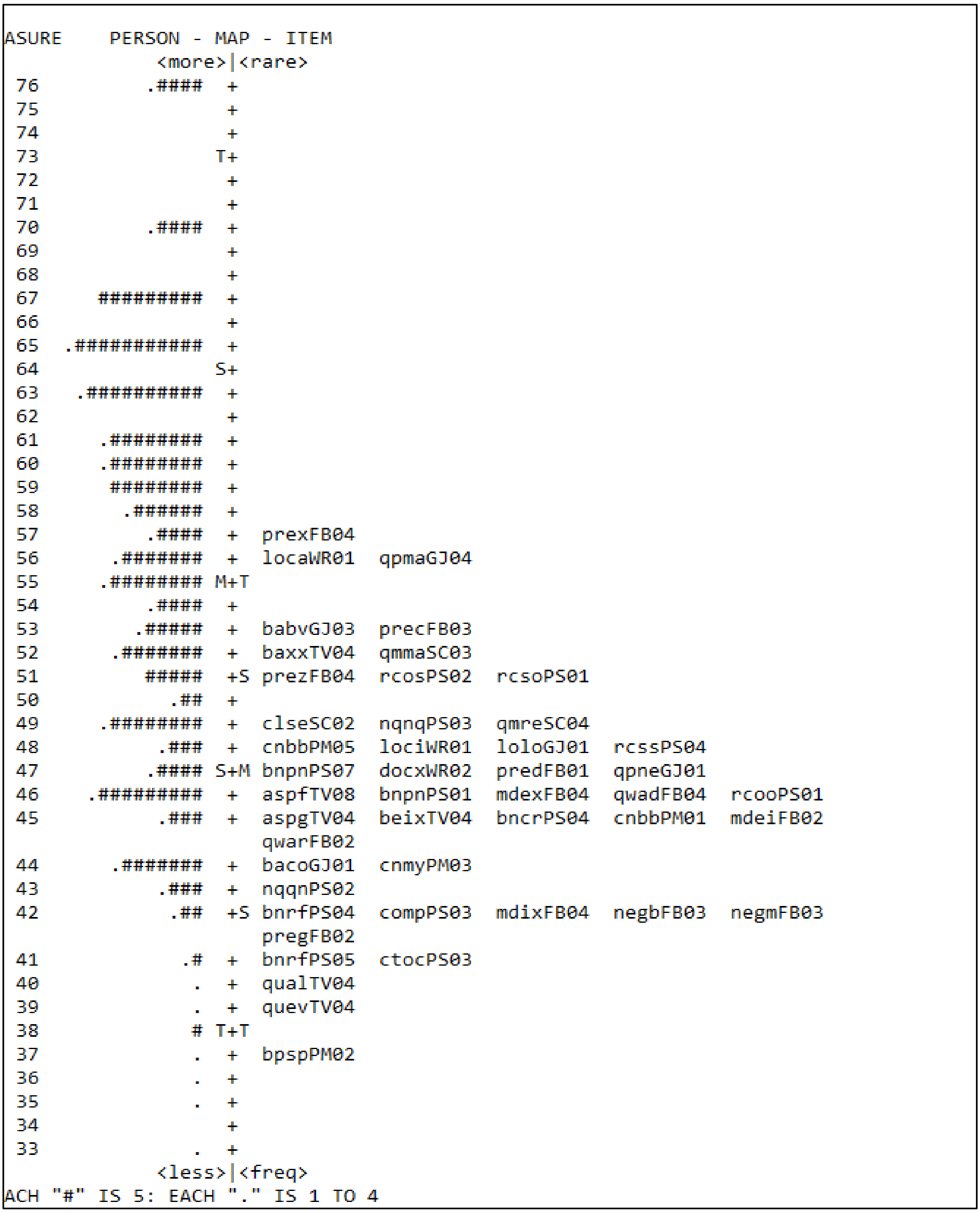
7.3.3.2 Evidence from Rasch Analysis

Reliability is a necessary condition for construct validity. The results of person separation reliability >.80 as estimated by Rasch analysis as well as a Cronbach alpha >.90 indicated good internal consistency of the items, reflecting that the items of the two lists of CGA were highly correlated with each other and converge to the same latent traits.

As discussed in Efeotor (2014), Wright maps are good evidence for our observation about the construct validity of the assessments. Figure 11 and Figure 12 show the two Wright maps for CGA-A and CGA-B respectively. The mean person ability of both forms showed to be higher than that of the mean item difficulty. The test items of both CGA-A and CGA-B were relatively easy for the typically developing participants in the study. More difficult items should be included to test for students with higher ability.

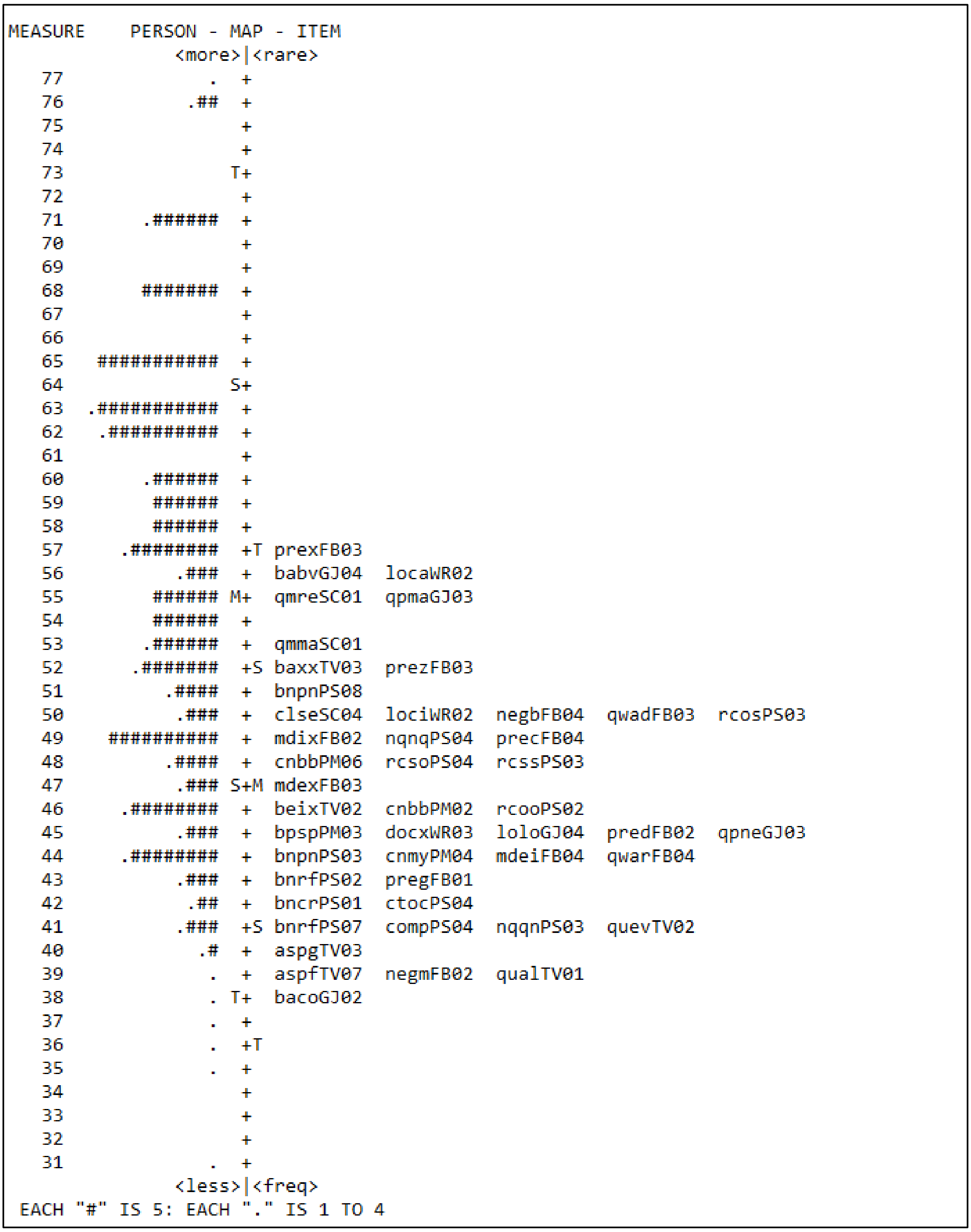
To further review the validity of the two lists, the two Wright maps for CGA-A (see Figure 13) and CGA-B (see Figure 14) based on DHH subjects were also generated for comparisons. The results were similar to the results based on the TD subjects, with a higher mean person ability than the mean item difficulty but the difference between overall person ability and item difficulty were relatively small, compared to the results from TD subjects. In sum, the test items of both CGA-A and CGA-B may be relatively easy for the tested group of participants.

When we observe the differences between the range of person abilities and the range of item difficulties based on the two Wright maps, the difficulty levels of the items are within the range of the persons' ability, only that the range of person ability is wider than the range of item difficulty. The differences between the two measures were around 30 logits. Even the most difficult items could not reach the level of the persons with higher abilities, that means, the two newly established CGA short tests may not be sensitive enough to discriminate the high ability group from the other participants. The results echo with that of the known-group validity reported in Section 7.3.2. As the data include students from P1 to P6, further investigation is required to see if CGA is reliable for senior primary school students.



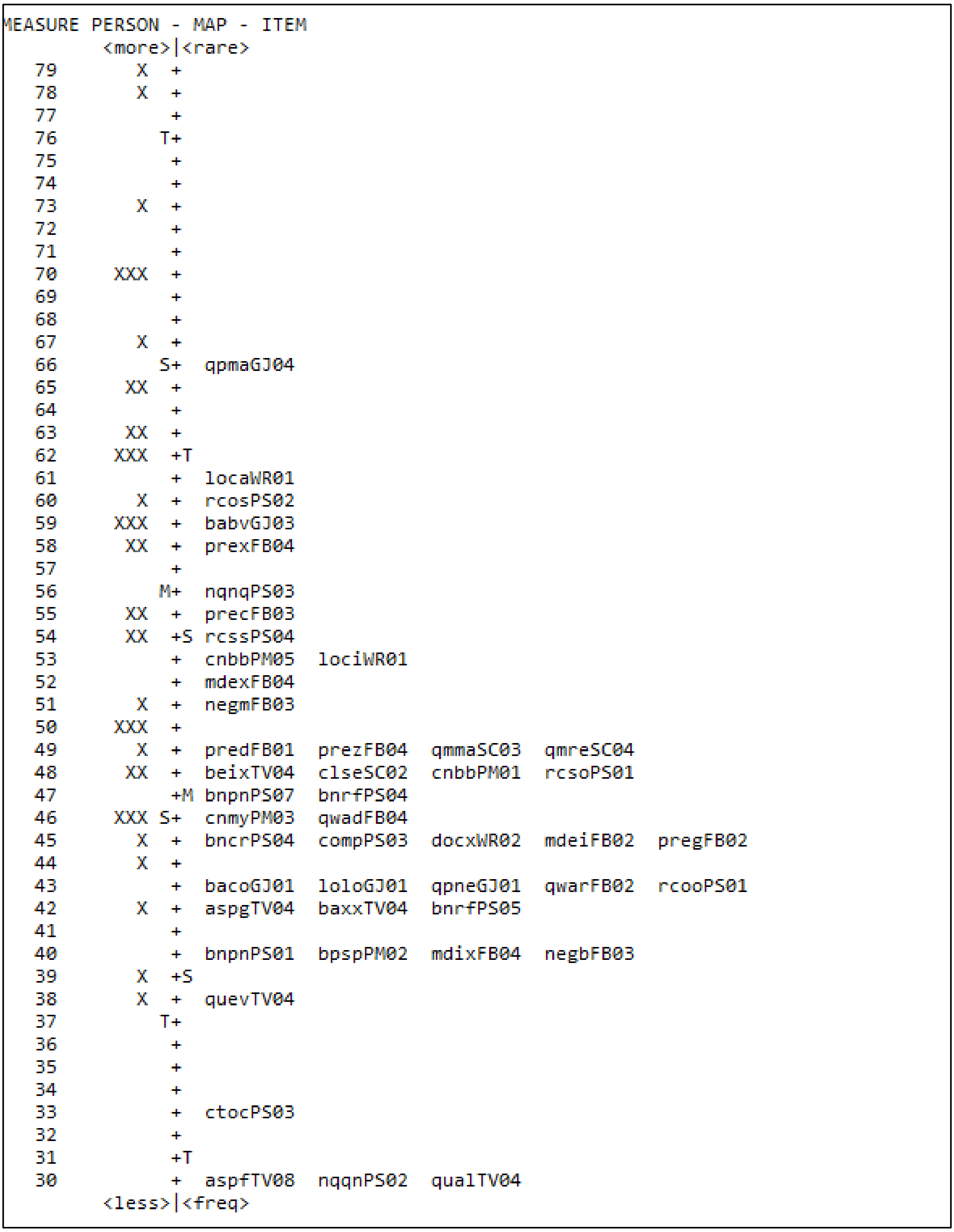
```
ASURE    PERSON - MAP - ITEM
             <more>|<rare>
 76         .####  +
 75                +
 74                +
 73               T+
 72                +
 71                +
 70         .####  +
 69                +
 68                +
 67     #########  +
 66                +
 65  .###########  +
 64               S+
 63   .##########  +
 62                +
 61     .########  +
 60     .########  +
 59      ########  +
 58       .######  +
 57         .####  +  prexFB04
 56      .#######  +  locaWR01  qpmaGJ04
 55     .######## M+T
 54         .####  +
 53        .#####  +  babvGJ03  precFB03
 52      .#######  +  baxxTV04  qmmaSC03
 51         #####  +S prezFB04  rcosPS02  rcsoPS01
 50           .##  +
 49     .########  +  clseSC02  nqnqPS03  qmreSC04
 48          .###  +  cnbbPM05  lociWR01  loloGJ01  rcssPS04
 47         .#### S+M bnpnPS07  docxWR02  predFB01  qpneGJ01
 46    .#########  +  aspfTV08  bnpnPS01  mdexFB04  qwadFB04  rcooPS01
 45          .###  +  aspgTV04  beixTV04  bncrPS04  cnbbPM01  mdeiFB02
                      qwarFB02
 44      .#######  +  bacoGJ01  cnmyPM03
 43          .###  +  nqqnPS02
 42           .##  +S bnrfPS04  compPS03  mdixFB04  negbFB03  negmFB03
                      pregFB02
 41            .#  +  bnrfPS05  ctocPS03
 40             .  +  qualTV04
 39             .  +  quevTV04
 38             # T+T
 37             .  +  bpspPM02
 36             .  +
 35             .  +
 34                +
 33             .  +
             <less>|<freq>
ACH "#" IS 5: EACH "." IS 1 TO 4
```

Figure 11. Wright map of CGA-A (831 TD subjects)



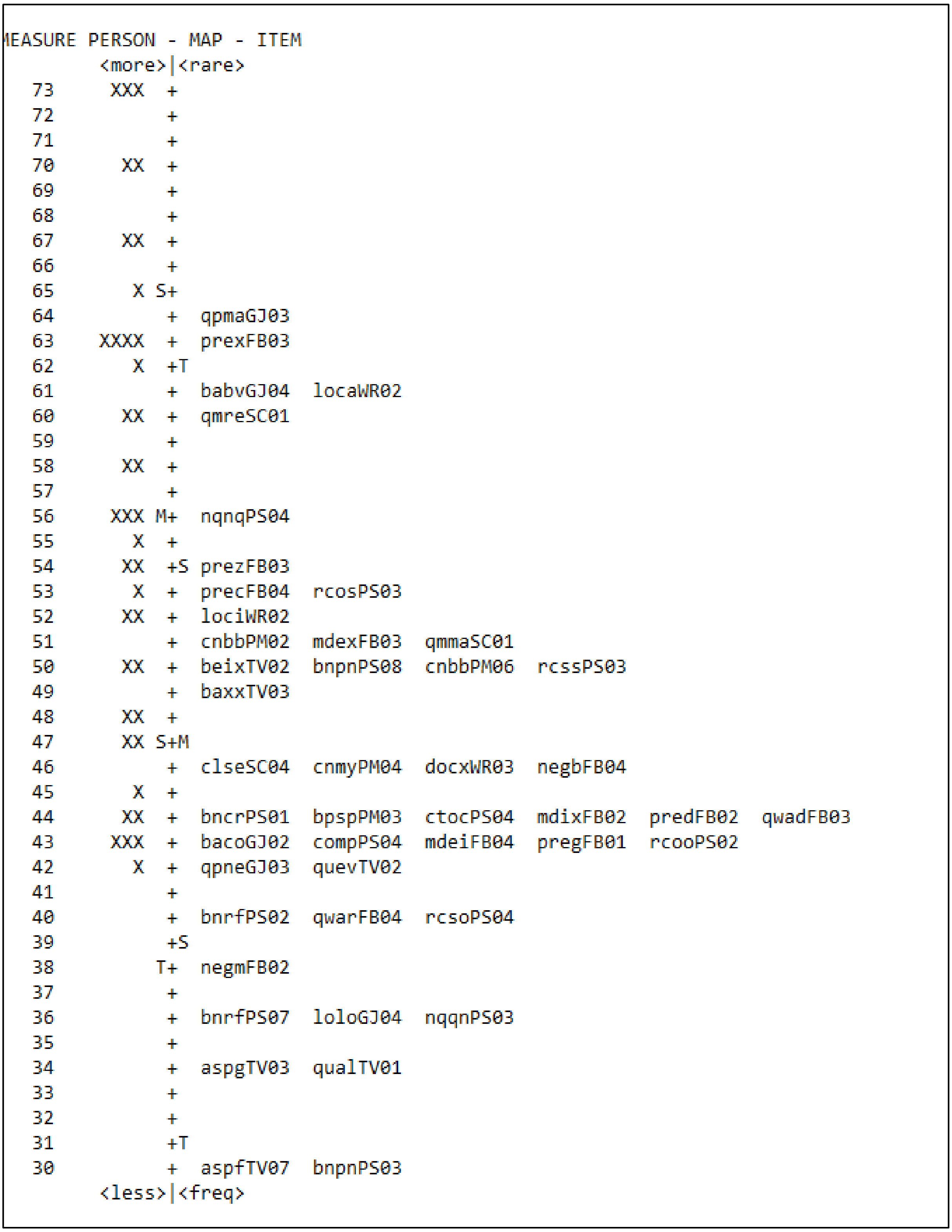
```
MEASURE    PERSON - MAP - ITEM
               <more>|<rare>
   77             .  +
   76           .##  +
   75                +
   74                +
   73               T+
   72                +
   71       .######  +
   70                +
   69                +
   68       #######  +
   67                +
   66                +
   65   ###########  +
   64               S+
   63  .###########  +
   62   .##########  +
   61                +
   60       .######  +
   59        ######  +
   58        ######  +
   57     .########  +T prexFB03
   56          .###  +  babvGJ04  locaWR02
   55        ###### M+  qmreSC01  qpmaGJ03
   54        ######  +
   53       .######  +  qmmaSC01
   52      .#######  +S baxxTV03  prezFB03
   51         .####  +  bnpnPS08
   50          .###  +  clseSC04  lociWR02  negbFB04  qwadFB03  rcosPS03
   49    ##########  +  mdixFB02  nqnqPS04  precFB04
   48         .####  +  cnbbPM06  rcsoPS04  rcssPS03
   47          .### S+M mdexFB03
   46     .########  +  beixTV02  cnbbPM02  rcooPS02
   45          .###  +  bpspPM03  docxWR03  loloGJ04  predFB02  qpneGJ03
   44     .########  +  bnpnPS03  cnmyPM04  mdeiFB04  qwarFB04
   43          .###  +  bnrfPS02  pregFB01
   42           .##  +  bncrPS01  ctocPS04
   41          .###  +S bnrfPS07  compPS04  nqqnPS03  quevTV02
   40            .#  +  aspgTV03
   39             .  +  aspfTV07  negmFB02  qualTV01
   38             . T+  bacoGJ02
   37             .  +
   36             .  +T
   35             .  +
   34                +
   33                +
   32                +
   31             .  +
               <less>|<freq>
 EACH "#" IS 5: EACH "." IS 1 TO 4
```

Figure 12. Wright map of CGA-B (831 TD subjects)



```
MEASURE PERSON - MAP - ITEM
         <more>|<rare>
   79       X  +
   78       X  +
   77          +
   76         T+
   75          +
   74          +
   73       X  +
   72          +
   71          +
   70     XXX  +
   69          +
   68          +
   67       X  +
   66         S+  qpmaGJ04
   65      XX  +
   64          +
   63      XX  +
   62     XXX  +T
   61          +  locaWR01
   60       X  +  rcosPS02
   59     XXX  +  babvGJ03
   58      XX  +  prexFB04
   57          +
   56         M+  nqnqPS03
   55      XX  +  precFB03
   54      XX  +S rcssPS04
   53          +  cnbbPM05  lociWR01
   52          +  mdexFB04
   51       X  +  negmFB03
   50     XXX  +
   49       X  +  predFB01  prezFB04  qmmaSC03  qmreSC04
   48      XX  +  beixTV04  clseSC02  cnbbPM01  rcsoPS01
   47          +M bnpnPS07  bnrfPS04
   46     XXX S+  cnmyPM03  qwadFB04
   45       X  +  bncrPS04  compPS03  docxWR02  mdeiFB02  pregFB02
   44       X  +
   43          +  bacoGJ01  loloGJ01  qpneGJ01  qwarFB02  rcooPS01
   42       X  +  aspgTV04  baxxTV04  bnrfPS05
   41          +
   40          +  bnpnPS01  bpspPM02  mdixFB04  negbFB03
   39       X  +S
   38       X  +  quevTV04
   37         T+
   36          +
   35          +
   34          +
   33          +  ctocPS03
   32          +
   31          +T
   30          +  aspfTV08  nqqnPS02  qualTV04
         <less>|<freq>
```

Figure 13. Wright map of CGA-A (39 DHH subjects)



```
MEASURE PERSON - MAP - ITEM
         <more>|<rare>
   73      XXX  +
   72           +
   71           +
   70       XX  +
   69           +
   68           +
   67       XX  +
   66           +
   65        X S+
   64           +  qpmaGJ03
   63     XXXX  +  prexFB03
   62        X  +T
   61           +  babvGJ04  locaWR02
   60       XX  +  qmreSC01
   59           +
   58       XX  +
   57           +
   56      XXX M+  nqnqPS04
   55        X  +
   54       XX  +S prezFB03
   53        X  +  precFB04  rcosPS03
   52       XX  +  lociWR02
   51           +  cnbbPM02  mdexFB03  qmmaSC01
   50       XX  +  beixTV02  bnpnPS08  cnbbPM06  rcssPS03
   49           +  baxxTV03
   48       XX  +
   47       XX S+M
   46           +  clseSC04  cnmyPM04  docxWR03  negbFB04
   45        X  +
   44       XX  +  bncrPS01  bpspPM03  ctocPS04  mdixFB02  predFB02  qwadFB03
   43      XXX  +  bacoGJ02  compPS04  mdeiFB04  pregFB01  rcooPS02
   42        X  +  qpneGJ03  quevTV02
   41           +
   40           +  bnrfPS02  qwarFB04  rcsoPS04
   39           +S
   38          T+  negmFB02
   37           +
   36           +  bnrfPS07  loloGJ04  nqqnPS03
   35           +
   34           +  aspgTV03  qualTV01
   33           +
   32           +
   31           +T
   30           +  aspfTV07  bnpnPS03
         <less>|<freq>
```

Figure 14. Wright map of CGA-B (39 DHH subjects)

7.3.3.3 Separate Analyses for TD and DHH Data

Table 38 shows the results of separately Rasch analysis for the two alternate forms, based on both TD and DHH subjects. The mean estimates of the person ability projected by the two alternate forms were found to have no big difference to each other (CGA-A: *M*=56.91 and CGA-B: *M*=56.44) based on DHH subjects. The mean estimates of the person ability were also found to have no great difference between CGA-A (*M*=55.49) and CGA-B (*M*=55.40) based on TD data. The standard deviations of DHH students' person abilities were large (CGA-A: *SD*=10.79 and CGA-B: *SD*=10.22) relative to those of the TD subjects (CGA-A: *SD*=8.69 and CGA-B: *SD*=8.92).

Table 38. Results of Separate Rasch Analysis for DHH (*N*=39) and TD subjects (*N*=831)

	CGA-A		CGA-B	
	Person	Item	Person	Item
A) DHH Subjects:	(*N*=39)	(*N*=46)	(*N*=39)	(*N*=46)
Range of logits	38.21-87.57	29.94-66.16	41.86-87.61	15.65-64.18
Mean	56.91	46.53	56.44	45.86
SD	10.79	7.88	10.22	9.04
Separation	2.51	2.02	2.43	2.09
Reliability	0.86	0.80	0.86	0.81
	Person	Item	Person	Item
B) TD Subjects:	(*N*=831)	(*N*=46)	(*N*=831)	(*N*=46)
Range of logits	33.05-84.79	36.69-56.83	31.42-85.16	38.25-57.45
Mean	55.49	46.53	55.40	46.53
SD	8.69	4.44	8.92	5.04
Separation	2.44	6.62	2.49	7.49
Reliability	0.86	0.98	0.86	0.98

When the item difficulty was considered, the means and standard deviations of the two alternate forms were quite similar (CGA-A: *M*=46.53 logits, *SD*=7.88; and CGA-B: *M*=45.88, *SD*=9.61) based on DHH data. The ranges of measures projected by the items of CGA-A and CGA-B were from 29.94-66.16 and from 15.65-64.18 respectively. The item with the lowest difficulty measure in CGA-B was aspfTV07 (with the item difficulty of 15.65 logit), which belongs to the grammatical category of "Aspect". CGA-B based on the DHH subjects showed to have a slightly larger range of difficulties when compared to that of CGA-A, but this phenomenon did not exist according to the results based on the TD subjects (see Table 38). The means and of the two alternate forms were the same with *M*=46.53 logits. Their standard deviations were slightly different from each other (CGA-A: *SD*=4.44; and CGA-B: *SD*=5.04) based on TD data, and the ranges of measures projected by the items of CGA-A and CGA-B were very similar, from 36.69-56.83 and from 38.25-57.45respectively.

When the two short forms were reviewed by their person reliability, both CGA-A and CGA-B were .86 (i.e., > .80) for both TD and DHH subjects. Their values of person separation for them were all >2 (from 2.43-2.51). The results indicate that the two alternate forms are able to discriminate the DHH students with different levels of ability regarding their grammatical knowledge in written Chinese. Reviewing the item reliability of the two alternate forms, their results of item reliability were .80 and .81 respectively (both are <.80) for DHH subjects while the results were very positive for TD subjects (reliability=.98 for both CGA-A and CGA-B). Moreover, the values of item separation were all <2 respectively for CGA-A and CGA-B, based on both DHH (2.02 and 2.09 respectively) and TD data (6.62 and 7.49 respectively). According to Linacre (1995), the results indicate a fair discrimination ability of CGA on persons' abilities, and this may be caused by a small DHH samples included for the analysis.

As suggested by Boone, Staver and Yale (2014) and Bond and Fox (2015), three types of Rasch measures should be considered for item fitness, including Outfit Mean Square Values (MNSQ), Outfit Z-Standardized Values (ZSTD), and Point Measure Correlation (PTMEA-CORR). The results of Outfit MNSQ can inform the researcher about the suitability of the item in measuring the validity, while PTMEA-CORR informs the extent to which the development of the constructs has achieved its goals (Bond & Fox, 2007). A positive PTMEA-CORR value indicates that the item measured the construct to be measured, while a negative PTMEA-CORR value indicates the opposite. On the other hand, ZSTD are t-tests for the hypothesis which can inform the researcher whether the data perfectly fits the model. Any item that fails to fulfill these three criteria needs to be improved or modified to ensure the quality and suitability of the item (Sumintono & Widhiarso, 2015). The recommended range for Outfit Mean Square Values (MNSQ) is 0.5-1.5, Outfit Z-Standardized Values (ZSTD) is -2.0-2.0, and Point Measure Correlation (PTMEA-CORR) is .40-.85 (Boone, Staver, & Yale, 2014). Only the items that fail to fulfill all the three criteria are required to be modified or deleted (Abul Aziz et al., 2014).

The items included in the two alternate forms were assessed thoroughly with their outfit MNSQ. All items had an outfit MNSQ out of the acceptable range had been excluded from the two forms. In addition, according to the item measures, all the items had a positive Point Measure Correlation, showing the items in the two CGA alternate forms are measuring for the same latent trait. Though some items had a ZSTD value out of the range of -2.0-2.0, no items had to be further removed according to the criteria suggested by (Sumintono & Widhiarso, 2015). The two alternate forms showed to be positively validated, confirming that the items of the two CGA short tests effectively measured the expected test construct.

7.3.4 Evidence from Dimensionality Measure

Other than item fitness, it is important to evaluate an instrument's dimensionality to ensure whether the test measures what it is supposed to measure (Abdul Aziz, Jusoh, Omar, Amlus, & Awang Salleh, 2014; Sumintono & Widhiarso, 2015). A single dimension was expected from both CGA-A and CGA-B. As discussed in Section 3.3.4, the criteria used for assessing the dimensionality of the two short forms stated in Table 10, following the suggestion from Sumintono and Widhiarso (2015, cited by Saidi and Siew, 2019, p.544). As an acceptable result, it is expected that the value of explained variance should be at least >20% and the unexplained variance in all other contrast should be ≤ 15%.

In this study, the dimensionality of the two alternate forms of CGA were 24.3% for CGA-A and 22.6% for CGA-B, both were above 20%. The largest unexplained contrast for the two lists were 9.4% for CGA-A and 9.3% for CGA-B, which were both smaller than 15%. According to the set criteria, the assumption that the two CGA forms are unidimensional was acceptable, though we could not get a very prominent result in this study. The result echoes with the analysis made by Min and Aryadoust (2021), which suggested that most assessments for grammatical knowledge were considered unidimensional in nature though the tasks might be different. In fact, the different tasks used in the assessment with different number of response choices may affect the results of the dimensionality measures.

In sum, according to the results of the different measures for CGA-A and CGA-B, the two short forms were found to have good reliability and validity. In addition, the equivalence of the two short forms were also confirmed. Further validation analyses would be continued to collect more evidence to prove the reliability and validity of the two short tests, especially when a new set of data was collected through the application of the two newly established short tests. Some additional measures, like test-retest reliability of the two tests, and the convergent validity between CGA and academic scores was also explored based on the new dataset with both TD and DHH subjects. The results will be reported in the following chapter.

Chapter 8: Validation of CGA with Newly Collected Data

The results of prior review of the reliability and validity of the two alternate forms was reported in Chapter 7. The validation was conducted based on the dataset extracted from the project "Profiling Chinese Grammatical Knowledge of Deaf and Hard-of-Hearing Students in HK and China – A Comparative Study" set up by the Centre for Sign Linguistics and Deaf Studies, The Chinese University of Hong Kong (Tang, et al., 2020). The subjects from the project were tested by different reliability and validity measures based on the items selected from the initial 172-item CGA profiling tool with no repeated measures. After developing the two equivalent lists of CGA items, the alternate forms reliability was conducted. The results were positive, showing that the two 46-item alternate forms were well validated with their psychometric properties. With the support of the evidence, two CGA short tests, namely CGA-A and CGA-B were established and used for field testing.

In this phase of study, the focus is to validate the two short tests with the data collected from a new group of subjects in 2022, before their norms were established for genuine education or clinical contexts to assess primary school students' grammatical knowledge in Chinese. To achieve this, it is important to apply the two short tests on a new group of subjects and further review their psychometric properties.

A new group of subjects were tested with both CGA-A and CGA-B following the standard procedures to review the alternate forms reliability between the two CGA short tests. A repeated measure using the same short tests was also conducted for establishing the test-retest reliability of CGA. The known-group validity was reviewed to see how CGA performance relates to grade levels of the students. Therefore, academic results in terms of reading and writing abilities of individual students were used to review the relationship between CGA and Chinese examination scores. In addition to the results of students' school examination results in Chinese Language, a set of normative data was also collected from the DHH students who were assessed by a standardized academic assessment in Chinese Language called the Learning Achievement Measurement Kit (LAMK; Education Bureau, 2008, 2014). This helps to establish the criterion validity of the two short forms of CGA. It is hypothesized that data from Chinese language examination is correlated with CGA test scores.

8.1 Instrument and Data Collection

The two 46-item short tests developed in this study, namely CGA-A and CGA-B, were used to assess students' Chinese grammatical knowledge. Before receiving the two short tests, all students were asked to complete a vocabulary pre-test to check if they understood the major vocabulary used in CGA. Besides collecting the assessment data from the two CGA short tests, data collection was also conducted with the consent of the school. It mainly includes students' examination scores of their Chinese language examination. The scores collected include two parts of students' scores: reading comprehension and writing. The scores were used to check for criterion validity of CGA. It is established in the literatures that students' grammatical knowledge in a language is highly correlated with their literacy skills (Kelly, 1996).

8.2 Participants

As mentioned in Chapter 3, this phase of study involves a group of TD and DHH students in a local primary school adopted a special programme called "Sign Bilingualism and Co-enrollment in Deaf Education Programme" (hereafter "SLCO Programme"), in which a critical mass of DHH students (like 4-5 DHH students in one co-enrollment class) were co-enrolled with TD students full-time in the same class (Yiu, Tang & Ho, 2019). The students were co-taught by a regular hearing teacher with a Deaf teacher or a hearing bilingual teacher who is competent in sign language skills. Both sign language and spoken language are used as the medium of instructions in all lessons.

In this case study, the TD and DHH students involved were studying in the same school. For each co-enrollment class with both TD and DHH students, students learn from the same group of teachers, following the same set of curricula in Chinese Language. Their results in CGA and academic performance can thus be compared with less confounding variables like the curriculum used for both TD and DHH students, the teachers' pedagogy, and other interventions such as reading programmes. The dataset collected can support some validity and reliability measures like test-retest or alternate forms reliability, however, the sample might be too homogenous that the results based on this group of students are not generalizable to the other group of TD or DHH students.

27 DHH and 112 TD students from P1 to P6 of the SLCO Programme were included for this case study. The distribution of the students by grade levels and their use of hearing devices were summarized in Table 11and Table 12 for reference.

8.3 Procedures

The TD students included in the assessment were tested in their classrooms together using iPad while the DHH students were tested individually in case signed communication is required. The two short tests were conducted online and so a set of iPad was used to assess the students individually. It took about 15-20 minutes for a student to complete one test. Teachers with proficient signing skills helped to conduct the assessment for DHH students. All students were instructed by a video showing how to answer questions in different tasks. They were then tested with the vocabulary pre-test, and then the two short tests, i.e., CGA-A and CGA-B. The order to prevent from learning effects during the assessment sessions, the test items of each CGA test were arranged by random. After students finished the first round of testing, they were asked to repeat the two short tests again within 1-3 weeks. The items in the re-test were presented with a different order, by conducting the test-retest reliability of CGA, "it is possible to estimate whether and to what extent the possible differences found on the measure are due to a real change in the person's ability or are within the measurement error of the test" (Holmefur et al., 2009, p.887).

After completing the two rounds of testing for the students, the data were checked thoroughly. Ten TD students were excluded from the dataset: among them, 3 students got vocabulary scores below 75%. 7 students had not completed all the tests (see Table 39 for the remaining data by grade levels). For DHH students, one P3 student got a vocabulary score of 72%, and one P4 student only completed the first round of test without completing the re-tests. Both data were kept for further review of CGA with a wider coverage of different grammatical knowledge in the data collection process.

Table 39. Number of TD and DHH data for the validity and reliability measures

	P1	P2	P3	P4	P5	P6	**Total**
TD students	12	15	17	22	19	17	**102**
DHH students	5	4	6	3	4	5	**27**

8.3.1 Descriptive Statistics

There are two rounds of assessments for one student, therefore two comparisons between CGA-A and CGA-B could be conducted for both TD and DHH students. The number “1” and “2” marked in the abbreviations such as CGA-A1 or CGA-A2 refers to the first or second round of CGA assessment for the students. Table 40 shows the descriptive statistics of the assessment results of the two CGA short tests for the two groups of students in raw scores. There was only a slight difference between the mean scores of CGA-A and CGA-B in both rounds of assessments. More detailed analyses of the reliability and validity of the assessment will be reported in the following sections.

Table 40. Mean and standard deviation of the test-retest results by CGA-A and CGA-B

Subjects	**Mean and SD of CGA Scores**			
	CGA-A1	**CGA-B1**	**CGA-A2**	**CGA-B2**
TD Subjects (*N*=102)	35.23 (8.15)	34.88 (7.62)	33.78 (9.27)	33.00 (9.21)
DHH Subjects (*N*=27)	33.48 (6.68)	32.70 (6.38)	33.27 (7.37)	32.69 (6.92)

*CGA-A1=1st round of CGA-A test results; CGA-A2=2nd round of CGA-A test results.
CGA-B1=1st round of CGA-B test results; CGA-B2=2nd round of CGA-B test results.

8.3.2 Reliability Measures

The results for the Alternate Forms Reliability and the Test-retest Reliability will be reported in the following sections. The Intraclass Correlation Coefficient (ICC) was used for the analysis. With reference to Koo & Li (2016), a single-measure, absolute-agreement, two-way mixed-effect model was used for the two analyses to check for the correlation between the two sets of data based on the analysis of data’s variance.

8.3.2.1 Alternate Forms Reliability between CGA-A and CGA-B

Table 41re shows the results of the alternate forms reliability between CGA-A and CGA-B in the first round (i.e. CGA-A1 and CGA-B1) and second round (i.e. CGA-A2 and CGA-B2) of assessment with a group of 102 typically developing (TD) and 27 deaf and hard-of-hearing (DHH) primary school students. The measures were conducted according to the raw scores of the students, and the analyses were made according to different groupings: TD only, DHH only and a combined group of both TD and DHH students. To investigate the alternate forms reliability and the test-retest reliability, Intra-class Correlation Coefficients (ICCs) was used for the analyses.

Table 41. Results of Intraclass Correlation Coefficients (ICC) for the evaluation of alternate forms reliability between CGA-A and CGA-B in two rounds of assessments

Alternate Forms Reliability#	Intraclass	95% Confidence Interval		F Test with True Value 0			
	Correlation	Lower Bound	Upper Bound	Value	*df* 1	*df* 2	Sig.
TD and DHH Subjects							
CGA-A1 vs. CGA-B1 (N=129)	.740	.651	.809	6.689	128	128	.000
CGA-A2 vs. CGA-B2 (N=128*)	.813	.745	.865	9.806	127	127	.000
TD Subjects							
CGA-A1 vs. CGA-B1 (N=102)	.718	.610	.801	6.071	101	101	.000
CGA-A2 vs. CGA-B2 (N=102)	.798	.715	.859	8.967	101	101	.000
DHH Subjects							
CGA-A1 vs. CGA-B1 (N=27)	.850	.700	.928	12.430	26	26	.000
CGA-A2 vs. CGA-B2 (N=26*)	.902	.831	.963	24.003	25	25	.000

# ICC estimates were calculated based on a single-measures, absolute-agreement, 2-way mixed-effects model

^TD Subjects =typically developing subjects; DHH Subjects= deaf and hrad-of-hearing subjects

**There was one missing data for the CGA-A2 and so only 26 subjects were included in this part of analysis*

The Intra-class Correlation Coefficients (ICCs) between CGA-A and CGA-B were .740 (95% *CI*=.651-.809) and .813 (95% *CI*=.745-.865) respectively for the first and second round of assessments, which indicate that the alternate forms reliability of CGA-A and CGA-B is positive when both TD and DHH subjects were involved (see Table 41). When the reliability coefficients were assessed separately for the raw scores of TD and DHH subjects, the results were .718 (95% *CI* =.610-.801) and .798 (95% *CI*=.715-.859) for the first and second round of assessments, i.e., CGA-A1 vs. CGA-B1, and CGA-A2 vs. CGA-B2 respectively for the TD subjects. For DHH subjects, the results were .850 (95% *CI* =.700-.928) and .902 (95% *CI*=.831-.963) for the first and second round of assessments respectively. In sum, the alternate forms reliability between CGA-A and CGA-B were “moderate to good” with reference to Koo and Li (2016). The results were better when the reliability coefficients were calculated based on the scores of the DHH subjects only. The range of ICCs was between .850 and .920, representing “good to excellent’ reliability.

In sum, the two alternate forms of CGA are considered reliable in testing the grammatical knowledge of written Chinese in both TD and DHH primary school students. The two CGA short tests can be used interchangeably for both TD and DHH students’ assessments, and the results projected from either forms of CGA are comparable with each other. According to the results, the two short tests can be used for tracking the developments of students by using the two tests alternatively.

8.3.2.2 Test-retest Reliability

Test-retest reliability is evaluated by testing subjects on repeated occasions. As mentioned that the TD and DHH students in this study were tested twice for both CGA-A and CGA-B. The aim is to check for stability of test results over time as one of the parameters to evaluate the reliability of the assessments (Holmefur et al., 2014). The Intra-class Correlation Coefficients (ICCs) were used based on a single-measure, absolute-agreement, two-way mixed-effects model. The results in different subject groups (TD only, DHH only and a combined group of TD and DHH subjects) were reported in terms of their raw scores (see Table 42)

Table 42. Results of Intraclass Correlation Coefficients (ICC) for the evaluation of test-retest reliability between the two rounds of assessments by both CGA-A and CGA-B

Test-retest Reliability#	**Intraclass**	**95% Confidence Interval**		**F Test with True Value 0**			
	Correlation	**Lower Bound**	**Upper Bound**	**Value**	***df* 1**	***df* 2**	**Sig.**
TD and DHH Subjects							
CGA-A1 vs. CGA-A2 (N=128*)	.727	.632	.800	6.513	127	127	.000
CGA-B1 vs. CGA-B2 (N=128*)	.630	.511	.725	4.549	128	128	.000
TD Subjects							
CGA-A1 vs. CGA-A2 (N=102)	.704	.589	.791	5.937	101	101	.000
CGA-B1 vs. CGA-B2 (N=102)	.586	.440	.701	3.977	101	101	.000
DHH Subjects							
CGA-A1 vs. CGA-A2 (N=26*)	.876	.744	.942	14.837	25	25	.000
CGA-B1 vs. CGA-B2 (N=26*)	.912	.817	.959	21.146	26	26	.000

# ICC estimates were calculated based on a single-measures, absolute-agreement, 2-way mixed-effects model
^TD Subjects =typically developing subjects; DHH Subjects= deaf and hrad-of-hearing subjects
**There was one missing data for the CGA-A2 and so only 26 subjects were included in this part of analysis*

The test-retest reliability between the two repeated measures of CGA-A (i.e. CGA-A1 and CGA-A2) and CGA-B (i.e. CGA-B1 and CGA-B2) were evaluated by calculating the Intraclass Correlation Coefficients (ICCs) between the repeated measures in raw scores. The results were also reported according to different groupings: TD only, DHH only and a combined group with both TD and DHH students.

According to the results summarized in Table 42, the intraclass coefficients for the repeated measures of CGA-A were .727 (95% *CI* =.632-.800) and the repeated measures for CGA-B were .630 (95% *CI* =.511-.725) based on the raw scores of both TD and DHH subjects. The test-retest reliability based on TD subjects was .704 (95% *CI*=.589-.791) for CGA-A and .586 (95% *CI* =.440-.701), representing a "moderate reliability". In contrast, the reliability was “good to excellent” based on the results of DHH subjects, which were .876 (95% *CI* =.744-.942) for the repeated measures of CGA-A and .912 (95% *CI*=.817-.959) for that of CGA-B. The results of test-retest reliability were not as good as expected. The two time points of the repeated measures may be too close to each other for some students. They may feel not interested in doing the test again within a short period of time and thus affect the performance. There exist some relatively poor results from a few TD subjects in their second round of assessments. Their scores were exceptionally low in their second round of assessments, for example, they got the raw scores of 40 (CGA-A) and 43 (CGA-B) in the first round of assessments, but only 17 (CGA-A) and 14 (CGA-B) in their second round of assessments.

8.3.3 Validity Measures

There were two areas of validity measures evaluated in this phase of study. Both the review of known-group validity and the convergent validity provided important evidence to support the validation of the two CGA short tests. The review of known-group validity focused on the relationship between CGA scores and the students’ grade levels. A significant main effect was expected.

For the measure of convergent validity, the focus is on the relationship between students’ Chinese grammatical knowledge and their academic performance in Chinese Language based on students’ school examination results. For the relationship between CGA and academic performance of DHH students, academic scores were also collected from normative assessment, namely the Learning Achievement Measuring Kit (LAMK; Education Bureau, 2008, 2014). A positive correlation >.80 is expected from the analysis.

8.3.3.1 Known-Group Validity

CGA as an assessment on Chinese grammatical knowledge, its development should have direct relationships with students' Chinese Language learning in school. Therefore, it is reasonable to expect that a longer time of Chinese Language learning in school should have better knowledge in Chinese grammatical knowledge. Under such an assumption, students at a higher grade level may obtain a higher score in CGA, no matter tested by CGA-A or CGA-B. In view that the sample included both TD (*N*=102) and DHH (*N*=27) students. Whether there was a main effect of hearing status or an interaction effect between hearing status and grade level should also be explored.

Levene's test suggested that the homogeneity of variance assumption was met. Two-way between-subjects ANOVA was conducted for the four different sets of CGA scores (namely CGA-A1, CGA-B1, CGA-A2, CGA-B2) from both TD and DHH subjects. Table 43 gives a summary of the results of two-way ANOVA, exploring the relationships between hearing status, grade levels and CGA scores. For all four test conditions reported in Table 43, there was no significant main effects of hearing status on CGA test scores (p >.05), and no significant interaction effect between grade level and hearing status (p >.05). The main effect of grade levels was significant with F (3,121)=5.618, p < 0.01 for CGA-A1, F(3,121)=7.378, p <.01 for CGA-A2, F(3,120)=3.210, p <.05 for CGA-B1, and F(3,121)=2.754, p <.05 for CGA-B2. The values of Partial Eta Squared ranged from .064 to .155, representing a medium to high effect size. In view that there was no significant effect of hearing status on CGA, TD and DHH subjects were grouped together for the following analyses.

Table 43. Results of two-way between-subjects ANOVA for evaluating the effects of grade levels and hearing status on the four rounds of CGA test scores (TD subjects, *N*=102; DHH, *N*=27)

Independent Variables	Test	Mean Square	df1	df2	F	Sig.	Partial Eta Squared
Grade Levels	CGA-A1	288.26	3	121	5.618**	.001	.122
	CGA-B1	317.05	3	121	7.378**	.000	.155
	CGA-A2	241.62	3	120^	3.210*	.026	.074
	CGA-B2	201.01	3	121	2.754*	.045	.064
Hearing Status	CGA-A1	1.53	1	121	0.030	.863	.000
	CGA-B1	2.15	1	121	0.050	.823	.000
	CGA-A2	3.62	1	120^	0.048	.872	.000
	CGA-B2	6.10	1	121	0.084	.773	.001
Hearing Status* Grade Levels	CGA-A1	82.46	3	121	1.607	.191	.038
	CGA-B1	45.95	3	121	1.069	.365	.026
	CGA-A2	17.67	3	120^	0.235	.872	.006
	CGA-B2	17.68	3	121	0.242	.867	.006

**p<0.05; **p<0.01*

^ *One DHH student did not take the second round of assessment by CGA-A.*

One-way ANOVA was conducted to investigate if there was a main effect of grade level on CGA scores. Table 44 gives a summary of the means and standard deviations of the CGA scores for all different grade levels. As discussed in Section 7.3.2, due to insignificant differences of CGA scores between grade levels from P4-P6, data from these three levels were combined to form a group for further analysis. An increasing trend of the students’ mean scores can be found from the results of all four test conditions. Statistical analysis confirmed that grade level was a significant factor on CGA scores for all four test conditions, with *p*< .01 (see Table 45). The results provides positive evidence to support known-group validity of CGA short tests.

Table 44. Mean and SD of CGA scores by grade levels (TD subjects, *N*=102; DHH subjects, *N*=27)

		1st Round of CGA Assessment				2nd Round of CGA Assessment		
	Grade Level	N	Mean	SD	Grade Level	N	Mean	SD
CGA-A	P1	17	26.82	8.263	P1	17	26.94	8.955
	P2	19	33.74	6.556	P2	19	33.37	7.127
	P3	23	35.00	5.901	P3	23	35.87	5.311
	P4-6	70	37.07	7.463	P4-6	69	34.70	9.600
	Total	129	34.86	7.876	Total	128	33.68	8.891
CGA-B	P1	17	26.29	6.980	P1	17	26.53	8.186
	P2	19	31.89	6.863	P2	19	31.42	8.624
	P3	23	35.43	4.962	P3	23	34.00	6.822
	P4-6	70	36.76	6.822	P4-6	70	34.47	8.887
	Total	129	34.43	7.411	Total	129	32.89	8.757

Table 45. Results of one-way ANOVA for assessing the main effects of grade levels on CGA in four test conditions (TD subjects, *N*=102; DHH subjects, *N*=27)

Subject	CGA Tests	*N*	*Mean (SD)*	*df1*	*df2*	*F*	Sig	Eta Squared
TD and DHH Subjects	CGA-A1	129	34.86 (7.88)	5	123	9.145**	.000	.366
	CGA-B1	129	34.43 (7.41)	5	123	9.908**	.000	.382
	CGA-A2	128	33.68 (8.89)	5	122	4.678**	.001	.250
	CGA-B2	129	32.89 (8.76)	5	123	5.216**	.000	.265

**p*<.05; ***p*<.01

Post-hoc Tukey tests were conducted to assess if there were significant mean differences between students' scores at different grade levels. Pairwise comparisons revealed that the main effect of grade level on CGA scores. As an overall review of the results in four test conditions, the effects were only significant between P1 and P3, and between P1 and P4-6, with *p*<.01 (see Table 46). No other significant mean difference could be clearly found between other grade levels.

Table 46. Post-hoc tests results regarding mean differences of CGA scores between different grade levels (TD subjects, *N*=102; DHH subjects, *N*=27)

		1st Round of CGA Assessment				2nd Round of CGA Assessment		
	Grade Level	P1	P2	P3	Grade Level	P1	P2	P3
CGA-A	P1				P1			
	P2	-6.91*			P2	-5.60		
	P3	-8.18**	-3.34		P3	-9.14**	-4.86*	
	P4-6	-10.25**	-1.26	-2.07	P4-6	-10.46**	-3.54	-1.32
CGA-B	P1				P1			
	P2	-6.43			P2	-4.89		
	P3	-8.93**	-2.50		P3	-7.47*	-2.58	
	P4-6	-7.75**	-1.33	-1.17	P4-6	-7.94**	-3.05	-0.47

**p*<.05; ***p*<.01

In view of the above results, known-group validity was further confirmed by the new set of data in view that the main effect of grade level on CGA scores was significant in all four test conditions, namely CGA-A1, CGA-A2, CGA-B1, and CGA-B2. Though the mean difference between adjacent grade levels was not statistically significant from P2 onward. The overall trend of better CGA scores obtained by students at higher grade levels was still confirmed. Pearson correlation revealed that grade levels and CGA scores were highly correlated with each other in all four CGA assessments, with the correlation coefficients *r* ranged from .30 to .49, with *p*<.01.

8.3.3.2 Convergent Validity

To check for convergent validity of CGA, one method is to examine correlations between CGA and some existing related measures for the latent trait, i.e., the Chinese grammatical knowledge of primary school students. The evaluation is to collect convergent evidence that supports the valid interpretation of assessment scores (Gioia, Espy, & Isquith, 2003). Because there was no standardized measure that was comparable to CGA, relationship between CGA and academic performance in Chinese Language education was thus explored.

Grammatical knowledge is a significant factor affecting reading development of both TD or DHH student data. Since there was no standardized comprehension assessment available for the study, students' examination scores in Chinese Language education (including reading comprehension and Chinese writing) were investigate if there is significant relationship between Chinese grammatical knowledge and the students' academic performance. In this study Chinese reading and writing examination scores were used to represent the academic performance of all TD and DHH subjects. In view that the examination papers were different for different grade levels, no fair comparison can be made between students from different grade levels. To facilitate further statistical analysis, the percentile ranks were calculated for individual grade levels. Therefore, for each grade, students' raw scores were used to rank the performance of individual students in terms of percentiles with reference to their same grade peers. Therefore, the performance of students was represented by their percentile ranks. A student with a better examination score was positioned at a higher percentile rank relative to his or her same grade peers.

Table 47. Correlation between CGA scores and students' percentile rank in Chinese reading and writing examinations (TD subjects, *N*=102; DHH subjects, *N*==27)

				Correlation	
Variables	*N*	Mean	*SD*	Chinese Reading	Chinese Writing
Chinese Reading	127	55.72	30.571		
Chinese Writing	127	54.04	29.455		
CGA-A1	128	34.89	7.899	.485**	.405**
CGA-B1	128	34.44	7.439	.390**	.383**
CGA-A2	127	33.68	8.926	.426**	.338**
CGA-B2	128	32.91	8.790	.325**	.241**

*p<.05; **p<.01

The results indicated that both CGA-A and CGA-B were positively correlated with the academic scores of the students in Chinese comprehension and writing (see Table 47). The correlation coefficients *r* ranged from .325 to .485, *p*<.01 for reading comprehension, and from .241 to .405, *p*<.01 for Chinese writing (see Table 47). Regression analysis was then conducted to see if students' CGA scores could significantly predict their academic performance in Chinese reading and writing. Results summarized in Table 48 showed that the regression model was significant with *p*<.01 in all the four test conditions. Students' performance in both CGA-A and CGA-B also significantly predicted their writing scores in their school examination (see Table 49).

Table 48. Regression analysis for the relationship between students' performance in CGA and Chinese reading (TD subjects, *N*=102; DHH subjects, *N*=27)

	ANOVA (Model Fitness)					**Regression Coefficients**		
Subject	*df*1	*df*2	*F*	Sig	Adjusted R^2	*B*	*t*	sig.
CGA-A1	1	124	16.03**	.000	.107	1.12	4.00**	.000
CGA-B1	1	125	7.71**	.006	.051	0.83	2.78**	.006
CGA-A2	1	124	27.55**	.000	.175	1.47	5.25**	.000
CGA-B2	1	125	14.74**	.000	.098	1.15	3.84**	.000

**p*<.05; ***p*<.01

Table 49. Regression analysis for the relationship between students' performance in CGA and Chinese writing (TD subjects, *N*=102; DHH subjects, *N*=27)

	ANOVA (Model Fitness)					**Regression Coefficients**		
Subject	*df*1	*df*2	*F*	Sig	Adjusted R^2	*B*	*t*	sig.
CGA-A1	1	125	24.53**	000	.157	1.54	4.95**	.000
CGA-B1	1	125	21.52**	.000	.140	1.51	4.64**	.000
CGA-A2	1	124	16.03**	.000	.107	1.12	4.00**	.000
CGA-B2	1	125	7.71**	.006	.051	0.83	2.78**	.006

**p*<.05; ***p*<.01

The review of convergent validity also included the measure of correlation between DHH students' Chinese grammatical knowledge and their performance in a normative academic achievement test, namely the Learning Achievement Measuring Kit (LAMK; Education Bureau, 2008, 2014). The LAMK was piloted in 2006, then revised and standardized in 2008 (Education Bureau, 2008). It is used to help identify and review the academic attainment of students with special education needs. All students with hearing loss had to receive the test in school to help review their academic progress to the Education Bureau. The Chinese Language assessment in LAMK has been validated in Rasch analysis against students with different grade levels in primary schools with excellent internal consistency (Cronbach's alphas .88).

The Chinese Language test paper of LAMK comprises of two sections, the reading and writing sections. It included reading comprehension of two stories and some writing tasks in 18 questions. The raw scores collected from students can be converted to standard scores with reference to the well-established norms. The test results help to define students' academic grade levels and to see if a student was achieving the same standard of his or her peers. They can also help to identify students with a delayed performance. The emphasis is to identify students with significant delay so as to alert teachers and clinicians of the current support for individualized learning support.

In this study, both the standard scores and the projected grade level were both used to reflect students' academic performance. The grade-level scores were converted as: "0=grade appropriate"; "1=one grade lower than student's current grade level"; "2=two grades lower than student's current grade level", etc. Therefore, a higher score represents a greater delay. Results showed that students' academic attainment as performance in standard scores of LAMK was positively correlated with their Chinese grammatical knowledge, represented by the CGA scores. The correlation coefficients r ranged from .738 to .838, p<.001, reflecting that there was a positive correlation between CGA and academic performance of DHH students (see Table 50). Results of linear regression supported CGA scores could significantly predict DHH students' academic attainment (in terms of the LAMK's standard scores of students), with a significant level of p<.01 (see Table 51). Therefore, no matter if students were tested by which CGA short test, the regression model was still significant with p<.01. The results reflected by the adjusted R^2 square values indicated that CGA test scores can explain 53%-69% of the total variance of the model. CGA scores can also predict students' academic standard in terms of grade levels. As mentioned above, the results of LAMK can help determine how many grade levels a student lag behind their same-age peers.

Table 50. Correlation between the CGA scores and the results of LAMK in terms of standard scores and projected grade levels (DHH subjects, N=26)

				Correlation	
Variables	*n*	Mean	*SD*	LAMK (Grade Level)	LAMK (Standard Score)
LAMK					
1.Grade Level	26	-0.65	1.16		
2.Standard Score	26	532.73	259.60		
CGA-A1	26	33.54	6.81	.622**	.808**
CGA-B1	26	32.62	6.49	.569**	.838**
CGA-A2	25	33.24	7.52	.627**	.738**
CGA-B2	26	32.46	7.01	.744**	.838**

*p<.05; **p<.01

Table 51. Regression analysis that investigates the relationship between students' performance in CGA and academic attainment in Chinese Language according to their standard scores in LAMK (DHH subjects, N=26)

	ANOVA (Model Fitness)					**Regression Coefficients**		
Subject	*df*1	*df*2	*F*	Sig	Adjusted R^2	*B*	*t*	sig.
CGA-A1	1	24	45.14**	.000	.638	30.81	6.72**	.000
CGA-B1	1	24	56.57**	.000	.690	33.50	7.52**	.000
CGA-A2	1	23	27.51**	.000	.525	25.70	5.25**	.000
CGA-B2	1	24	56.83**	.000	.691	31.05	7.54**	.000

*p<0.05; **p<0.01

Table 52. Regression analysis that investigates the relationship between students' performance in CGA and their academic performance in Chinese Language according to the grade levels projected by LAMK (DHH subjects, *N*=26)

	ANOVA (Model Fitness)					Regression Coefficients		
Subject	*df1*	*df2*	*F*	Sig	Adjusted R^2	*B*	*t*	sig.
CGA-A1	1	24	15.11**	.000	.361	.106	3.89**	.001
CGA-B1	1	24	11.47**	.002	.295	.102	3.39**	.002
CGA-A2	1	23	14.89**	.000	.367	.096	3.86**	.001
CGA-B2	1	24	29.80**	.000	.535	.124	5.46**	.000

**p*<.05; ***p*<.01

CGA scores (no matter tested by CGA-A or CGA-B) were negatively correlated with the years of delay in students' academic level in Chinese Language (see Table 50). This result indicated that the better the CGA performance, the lesser the delay in students' level of Chinese academic attainment. Results of linear regression also confirmed that CGA scores could significantly predict the level of delay in student's academic attainment, with *p*<.01 for all four test conditions (see Table 52). The total variance explained by the independent variable, that is CGA scores is from 30% to 54%.

The above-mentioned results further confirm the convergent validity of CGA and reiterate the significance of the development of CGA, which helps to review Cantonese-speaking TD and DHH students' grammatical knowledge in written Chinese. The assessment results can also bring attention to teachers the possible learning needs of students in Chinese Language, including their reading and writing abilities.

8.4 A Summary

After two rounds of reliability and validity measures, using the old data from the database provided by the Centre for Sign Linguistics and Deaf Studies, The Chinese University of Hong Kong (Tang et al., 2020), and the newly collected data in 2022. With two different sets of assessment data collected from both TD and DHH students, the CGA scores of the two alternate forms or short tests were assessed for their reliability and validity by different measures. The measures were complementary to each other to collect a more complete set of evidence to support the validation of the CGA (see Table 53). In sum, good reliability and validity were confirmed by the results of different measures based on the two sets of data. The norming procedures were thus proceeded, and the result is reported in the following chapter.

Table 53. A summary of the results of reliability and validity measures conducted for the two CGA short tests

Reliability and Validity Measures	Use of Dataset[a]	Results
1. Separation reliability	Dataset 2015-19	Good reliability >.80
2. Internal consistency	Dataset 2015-19	Good internal consistency >.90
3. Alternate forms reliability	Dataset 2015-19 & Dataset 2022	Moderate to excellent reliability (.75-.90)
4. Test-retest reliability	Dataset 2022	Moderate to good reliability (.50-.75)
5. Content validity	Dataset 2015-19	Good ratings >4.9 & CVIs >.90
6. Known-group validity	Dataset 2015-19 & Dataset 2022	Significant grade difference
7. Convergent validity	Dataset 2022	Significantly correlated with academic performance in Chinese Language
8. Construct validity	Dataset 2015-19	Good evidence supporting construct validity
9. Dimensionality	Dataset 2015-19	Acceptable as unidimensional

[a] "Dataset 2015-19" refers to the initial collected data from the CGA profiling project (Tang et al., 2020); "Dataset 2022" refers to the newly collected data from the Sign Bilingualism and Co-enrollment in Deaf education Programme.

Chapter 9: Developing the Norms of CGA

After a series of psychometric review and evaluation the validity and reliability of the two CGA short tests, the findings supported that the two alternate forms of CGA are valid and reliable in assessing typically developing (TD) and deaf or hard-of-hearing (DHH) students' Chinese grammatical knowledge. The two CGA short tests newly used in a group of TD and DHH students are also proved to have positive alternate forms reliability and test-retest reliability. For future applications in the educational and clinical settings, it would be conducive to establish the norms for the two short tests. Though the remaining dataset for each grade level was reduced in certain proportion after conducting different analyses, the dataset available has gone through a stringent review process, they are all considered good-fit items and persons with good reliability and fitness to the model. It would be a functional contribution to the practitioners if a norm can be set up for each of the two CGA short tests. To facilitate general practitioners' applications, the norms would be developed based on the raw data, rather than the estimated logits through Rasch analysis.

9.1 Establishing the Norms for Different Grade Levels

As discussed in the earlier section, the two CGA short tests are proved to be significant in discriminating students at different grade levels, however, both CGA-A and CGA-B are more effective in identifying the differences between students at the first three grade levels (from P1-P3) than the students at senior grade levels from P4 to P6 (see results in Section 7.3.2). This phenomenon is also observed from the descriptive statistics calculated through Rasch analysis such as mean, standard deviation and range of person ability (see Section 7.3.3) for the two CGA equivalent lists. The Wright maps (from Figure 11 to Figure 14) generated also provide evidence that the two CGA short tests lack sufficient items with higher difficulty to discriminate abilities of senior primary students or high-ability students. Though the one-way ANOVA conducted by the Welch's F-test has confirmed that student's grade level has a significant main effect on the CGA scores in logits, post hoc tests indicated that the mean differences between adjacent grade levels from P3 onward are not significant statistically (see Table 34 and Table 35). Students from P4-P6 performed like one group of students with similar person ability. Therefore, even though there are norms for individual grade levels of P4, P5 and P6, the standard among these three groups of students may be very close to each other. In another words, this group of students are not distinguishable from each other according to their CGA scores even though they are studying at different grade levels. A regrouping of the data from P4-P6 as a single grade level was thus proposed.

The "P4-6" group represents students from the three senior primary levels from P4 to P6. The norm for this group is established, aims to see if a student from P4 onward is performing within an acceptable range of performance in their Chinese grammatical knowledge relative to their peers from P4-P6. In this phase of study, a statistical analysis would be conducted to further verify the proposal of data regrouping in four grade levels using the CGA raw scores. The norms for CGA-A and CGA-B in four grade levels, namely "P1", "P2", "P3" and "P4-6", would be set up in terms of percentile ranks. Finally, some further analyses would be conducted to define the cutting points in terms of their percentile ranks that help to identify students with the needs of additional support for their development of Chinese grammatical knowledge.

The norms of the two short tests are set up in terms of TD students' raw scores, so it will be easier for educators and clinicians to check for students' results after testing them in either

CGA-A or CGA-B. The norms were established according to the raw scores of the 831 TD students collected from TD students at different local primary schools. One-way ANOVA were thus conducted again to further reviewed the proposed grouping of students based on the raw scores to review if there were significant grade difference between students' CGA scores. The mean difference between their performance at different grade levels were also explored.

9.1.1 Analyses for Grade Difference (Mean Plots)

The analysis was conducted for data in six grade level first. Figure 15 and Figure 16 are the mean plots, showing an increasing trend of the mean CGA scores by the six grade levels, but the slope of the plots turns mild after P3. According to the mean scores of TD students at different grade levels, the raw scores were increased with grade levels (see Table 54 and Table 55). The mean difference from P4 onward was smaller than those from P1 to P3 for both lists of CGA.

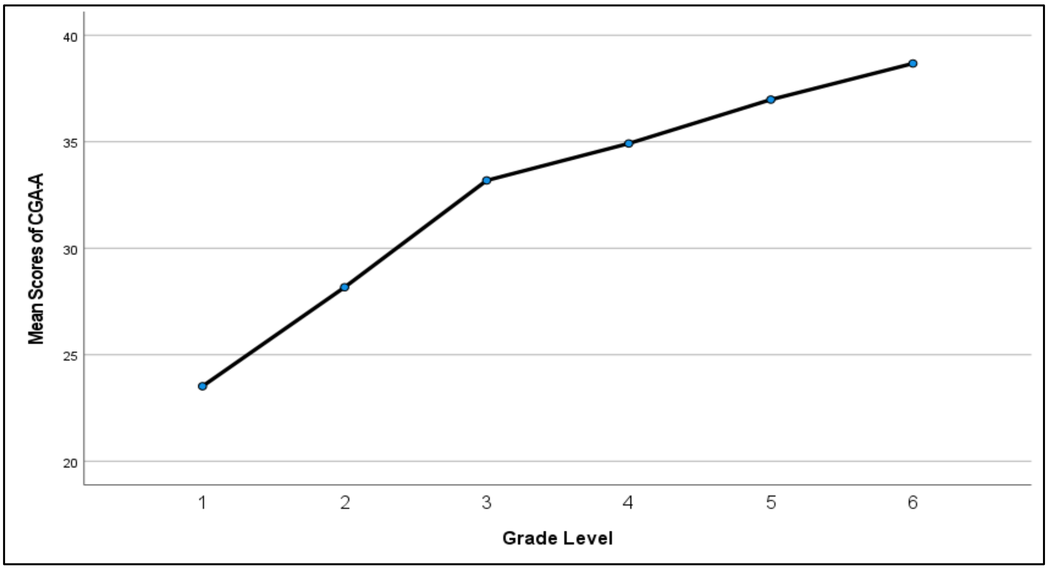


Figure 15. Mean plots for CGA-A by six grade levels in raw scores (TD subjects)

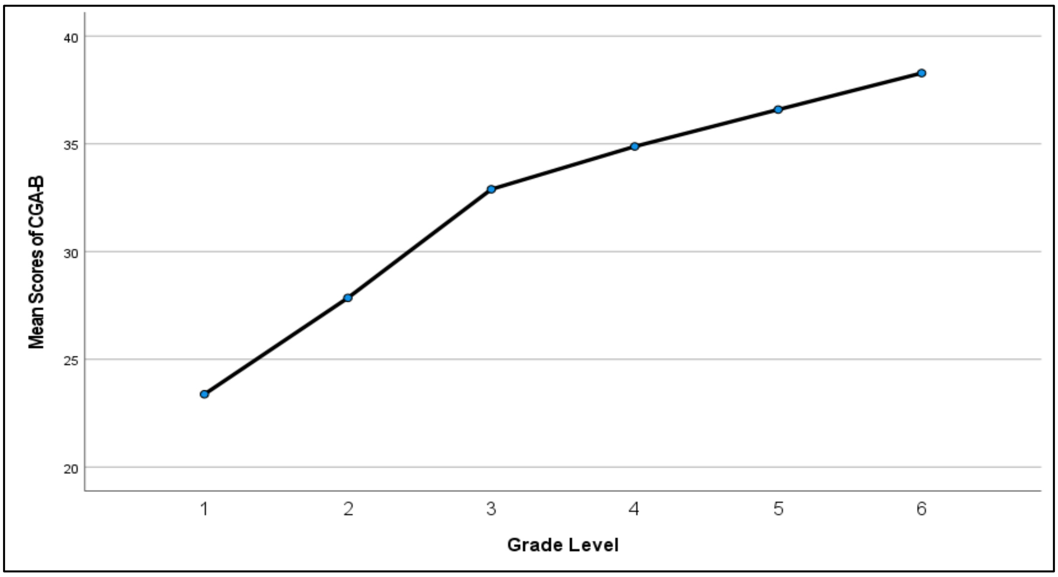


Figure 16. Mean plots for CGA-B by six grade levels in raw scores (TD subjects)

As discussed in Section 7.3.2, a 4-Grade model may be more suitable for the existing dataset. Therefore, besides students from P1, P2, and P3, the students from P4 to P6 were grouped together, representing a group of senior primary students. Figure 17 and Figure 18 represent the mean plots for CGA-A and CGA-B after the re-grouping. According to the figures, a more stable increase of CGA scores were observed in the 4-Grade model as compared to the mean plots in the 6-Grade model (see also Figure 15 and Figure 16).

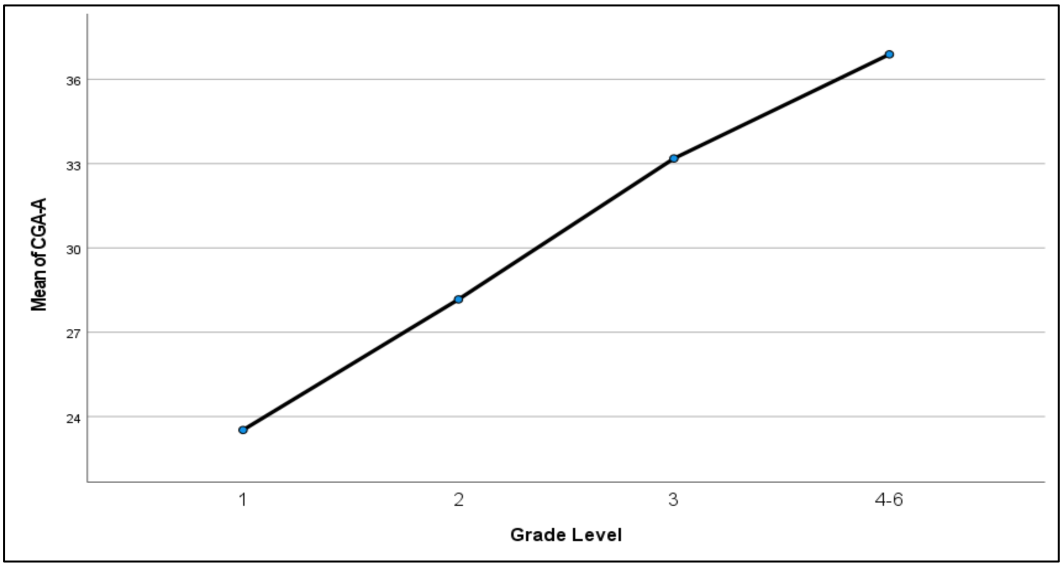


Figure 17. Mean plot for CGA-A grouped in four grade levels (TD subjects, *N*=831)

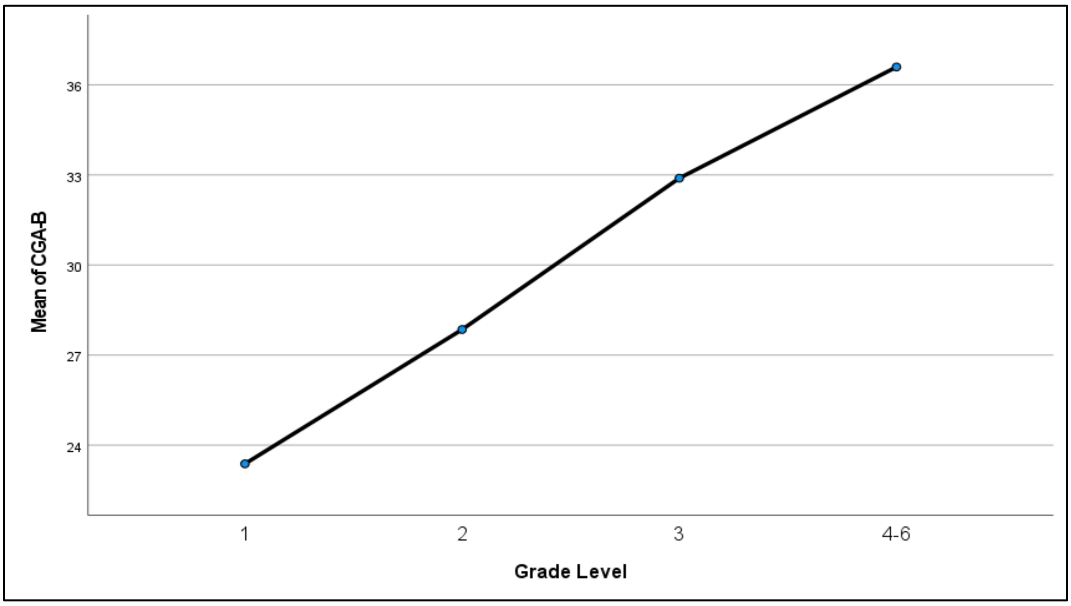


Figure 18. Box plot for CGA-B grouped in four grade levels (TD subjects, *N*=831)

9.1.2 Analyses for Grade Difference (One-way ANOVA)

After observing the mean plots, one-way ANOVA was conducted. By checking the data's normality using the Shapiro-Wilk test and the homogeneity of variance using the Levene's statistics, the assumption of normality and the homogeneity of variance were violated with a significant level of p<0.01 in both analyses, and the results apply to both students' groupings, i.e., grouping students from P1 to P6 in six grade levels (The 6-Grade Model) or grouping them in 4 grade levels, i.e., P1, P2, P3, and P4-6 (The 4-Grade model). Therefore, the analysis of variance for both models were conducted by Welch's *F*-test, and the subsequent pairwise comparisons of mean differences between different grade levels were conducted through the Games-Howell procedures.

9.1.2.1 The 6-Grade Model

The analysis was first conducted for the 6-Grade Model, in which students were grouped according to their six grade levels, i.e., P1, P2, P3, P4, P5, and P6. The results of Welch's *F*-test confirmed that there was a significant main effect of grade levels, with Welch $F(5, 324.43)=69.96$, $p<.001$, $\omega^2=0.32$ for CGA-A, and $F(5, 325.55)=66.60$, $p<.01$, $\omega^2=0.31$ for CGA-B (see Table 54 and Table 55 for the results of CGA-A and CGA-B respectively).

Table 54. Results of Post-hoc Games-Howell Test for grade differences of TD students based on both the 6-Grade Model and the 4-Grade Model tested by CGA-A (TD Subjects, *N*=831)

			Mean Difference (Raw Scores)				
Grade	*N*	Mean (*SD*)	P1	P2	P3	P4	P5
A) The 6-Grade Model[a]							
P1	84	23.38 (8.38)					
P2	170	27.85 (7.38)	-4.65**				
P3	216	32.89 (7.96)	-9.66**	-5.02**			
P4	100	34.88 (7.59)	-11.40**	-6.75**	**-1.74**		
P5	159	36.59 (6.92)	-13.46**	-8.81**	-3.80**	**-2.06**	
P6	102	38.28 (7.59)	-15.15**	-10.51**	-5.49**	-3.76**	**-1.70**
B) The 4-Grade Model[b]							
P1	84	23.38 (8.38)					
P2	170	27.85 (7.38)	-4.65**				
P3	216	32.89 (7.96)	-9.66**	-5.02**			
P4-6	361	36.60 (7.39)	-13.37**	-8.72**	-3.70**		

[a] The 6-Grade Model refers to students' grouping into 6 grade levels from P1 to P6.

[b] The 4-Grade Model refers to students' grouping into 4 grade levels, namely P1, P2, P3, and P4-6.

* *p*<.05, ** *p*<.01

The post-hoc Games-Howell tests results found that the mean differences of CGA scores were significant between P1-P2 (*p*<.01) and P2-P3 (*p*<.01) for both CGA-A and CGA-B, but were not significant between P3-P4, P4-P5 and P5-P6. Similar to the results in logit measures, significant mean difference could only be found between P3 and P5 (*p*<.001), and between P4 and P6 (*p*<.001). As showed in the mean plots (see Section 9.1.1) and the discussion made in Section 7.3.2, the 4-Grade Model may be more suitable for the development of the norms according to the existing norming data from TD students.

Table 55. Results of Post-hoc Games-Howell Test for grade differences of TD students based on both the 6-Grade Model and the 4-Grade Model tested by CGA-B (TD Subjects, *N*=831)

			Mean Difference (Raw Scores)				
Grade	N	Mean (SD)	P1	P2	P3	P4	P5
P1	84	23.38 (8.38)					
P2	170	27.85 (7.38)	-4.47**				
P3	216	32.89 (7.96)	-9.51**	-5.04**			
P4	100	34.88 (7.59)	-11.50**	-7.03**	**-1.99**		
P5	159	36.59 (6.92)	-13.21**	-8.74**	-3.70**	**-1.71**	
P6	102	38.28 (7.59)	-14.90**	-10.43**	-5.39**	-3.40**	**-1.69**
B) The 4-Grade Model[b]							
P1	84	23.38 (8.38)					
P2	170	27.85 (7.38)	-4.47**				
P3	216	32.89 (7.96)	-9.51**	-5.04**			
P4-6	361	36.60 (7.39)	-13.22**	-8.74**	-3.70**		

[a] The 6-Grade Model refers to students' grouping into 6 grade levels from P1 to P6.
[b] The 4-Grade Model refers to students' grouping into 4 grade levels, namely P1, P2, P3, and P4-6.
* *p*<.05, ** *p*<.01

9.1.2.2 The 4-Grade Model

To investigate if the 4-Grade Model is more suitable for the development of the norms for the two short tests, one-way ANOVA was conducted for this new grouping, and the results confirmed that grade level had a significant main effect on students' performance in CGA-A and CGA-B (see Table 54 and Table 55). The results for CGA-A was $F(5, 287.76)=94.61$, $p<.001$, $\omega^2=.31$ for CGA-A, and $F(5, 289.86)=92.10$, $p<.001$, $\omega^2=.30$ for CGA-B.

Post-hoc Games-Howell test results showed that there were significant mean difference between students at "P4-6" and all other three junior primary grade levels (i.e. P1, P2 and P3) with the significant levels of $p<.01$ for both CGA-A and CGA-B (see Table 54 Table 55 respectively). An estimate of 30-31% of the total variance (ω^2= .31 for CGA-A and .30 for CGA-B) of the dependent variable (i.e. student's raw scores in CGA) is accounted for by students' grade levels, as the independent variable.

With the above analyses, the norms for the two short versions of CGA were thus set up based on 4-Grade Model, namely P1, P2, P3, and P4-6. There is no specific norm set for individual grade levels from P4 to P6.

9.2 Establishing the Norms in Percentile Ranks

With the support of the positive results from different psychometric review conducted, the two short version of CGA as alternate forms are confirmed to have satisfactory reliability and validity through different validation procedures. CGA-A and CGA-B were thus developed as two normative assessments to measure primary school students' grammatical knowledge in written Chinese. The two short tests can be used as screening tests to help identify if a primary school student is lagging behind their peers based on the norms set for the four grade levels defined for CGA. Based on the norm established in percentile ranks, students who are observed to be delayed in their development in Chinese grammatical knowledge can be supported with respective interventions and reviewed for progress using the two alternate forms of CGA.

Table 56. The test of normality for the data of CGA in four grade levels (TD subjects, *N*=831)

Tests of Normality							
		Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Grade Level	Statistic	*df*	Sig.	Statistic	*df*	Sig.
CGA-A	P1	.084	84	**.200***	.977	84	**.141**
	P2	.092	170	.001	.979	170	.012
	P3	.136	216	.000	.929	216	.000
	P4-6	.182	361	.000	.870	361	.000
CGA-B	P1	.109	84	.015	.971	84	**.057**
	P2	.071	170	.035	.980	170	.016
	P3	.116	216	.000	.940	216	.000
	P4-6	.184	361	.000	.871	361	.000

*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction

By conducting the tests for normality, namely the Kolmogorov-Smirnov Test and the Shapiro-Wilk Test, it was confirmed that the data of CGA raw scores in 4 grade levels were not normally distributed for most grade levels, with the significant levels of *p*<.05 for grade levels of P2, P3 and P4-6 (see **Error! Reference source not found.**). The normality assumption was met only f or CGA scores of P1 students, with the results of .977 (*df*=84, *p*=.141) for CGA-A and .971 *(df*=84, *p*= .057) for CGA-B. In view that most of the data do not meet with the normality assumption, the norms for CGA-A and CGA-B were set up in terms of percentile ranks rather than standard deviation or *z*-scores.

9.2.1 Converting Raw Scores to Percentile Ranks

Table 57. Descriptive statistics for CGA-A and CGA-B (TD subjects, *N*=831)

Alternate Forms	Grade	*N*	*Mean*	*Median*	*SD*	95% Confidence Interval	
CGA-A	P1	84	23.52	23.00	8.42	21.70	25.35
	P2	170	28.17	28.00	7.39	27.05	29.29
	P3	216	33.19	35.00	7.82	32.14	34.23
	P4-P6	361	36.89	39.00	7.17	36.15	37.63
CGA-B	P1	84	23.38	22.00	8.38	21.56	25.20
	P2	170	27.85	27.00	7.38	26.74	28.97
	P3	216	32.89	34.00	7.96	31.83	33.96
	P4-P6	361	36.60	40.00	7.39	35.83	37.36

The raw scores of the two short tests range from 0-46. The means of all grade levels were within the 95% Confidence Intervals (see descriptive statistics in Table 57). For the development of the norms for the two short tests, the equivalent percentile ranks for each grade levels were calculated based on the raw scores collected from the 831 TD primary students using SPSS version 27. As showed in Table 58, the percentile ranks projected by the raw scores of CGA-A and CGA-B were very similar according to our surface observation. Further statistical review on the alternate forms reliability of the two short tests and the respective norms will be reviewed based on a new set of TD and DHH data.

With the conversion tables set for the norms of CGA-A and CGA-B by grade levels, whenever there are students assessed by the two CGA short tests, teachers or clinical practitioners can simply check with the conversion tables for their percentile ranks with reference to their grade levels. They can also compare the students' results when a re-test is conducted. CGA-A and CGA-B can be used inter-changeably to check for the students' development or to monitor specific progress after different literacy interventions. A crucial question that we need to answer is "below which percentile rank that a student should be considered as having a delayed development in CGA". At this stage, as a newly developed assessment tool, there is no evidence to help define the cut-off points for students with different levels of performance. Further evidence collected from empirical research will be essential.

Table 59 shows the percentile ranks for CGA-A and CGA-B by the four different grade levels. For example, a raw score of "25" tested by CGA-A has an equivalent percentile rank of "63" for a student at P1, but a percentile rank of "37" and "21" for students at the level of P2 and P3 respectively. As discussed above, students from P4-P6 are grouped together in terms of the norms for CGA. Therefore, the equivalent percentile ranks for a raw score of "25" are the same for all students studying at P4, P5 or P6. For example, if a student of P4 is tested by CGA-A, with a score of "25", his or her percentile rank will be "10", if he or she is tested by CGA-B with the same score of "25", his or her percentile rank will be "12".

Table 58. Percentile ranks projected by the equivalent raw scores of CGA-A and CGA-B according to the 4-Grade Model

	Percentile Ranks									
Grade Levels	5th		25th		50th		75th		95th	
	CGA-A	CGA-B	CGA-A	CGA-B	CGA-A	CGA-B	CGA-A	CGA-B	CGA-A	CGA-B
P1	10.25	10.25	19.00	17.00	23.00	22.00	30.75	30.75	37.75	38.75
P2	17.00	16.00	22.75	22.00	28.00	27.00	34.25	34.00	39.00	40.46
P3	19.00	17.85	27.00	27.00	35.00	34.00	40.00	40.00	43.00	43.00
P4-6	22.00	21.00	34.00	32.00	39.00	40.00	42.00	42.00	45.00	44.00

With the conversion tables set for the norms of CGA-A and CGA-B by grade levels, whenever there are students assessed by the two CGA short tests, teachers or clinical practitioners can simply check with the conversion tables for their percentile ranks with reference to their grade levels. They can also compare the students' results when a re-test is conducted. CGA-A and CGA-B can be used inter-changeably to check for the students' development or to monitor specific progress after different literacy interventions. A crucial question that we need to answer is "below which percentile rank that a student should be considered as having a delayed development in CGA". At this stage, as a newly developed assessment tool, there is no evidence to help define the cut-off points for students with different levels of performance. Further evidence collected from empirical research will be essential.

Table 59. Conversion table of equivalent percentile ranks for CGA-A and CGA-B based on the raw scores of the TD students (*N*=831)

Raw Score	Percentiles for CGA-A				Raw Score
	P1	P2	P3	P4-P6	
0	1	1	1	1	0
1	1	1	1	1	1
2	1	1	1	1	2
3	1	1	1	1	3
4	1	1	1	1	4
5	1	1	1	1	5
6	1	1	1	1	6
7	1	1	1	1	7
8	2	1	1	1	8
9	4	1	1	1	9
10	5	2	1	1	10
11	7	2	1	1	11
12	12	3	1	1	12
13	14	3	1	1	13
14	18	4	1	1	14
15	21	4	1	1	15
16	23	4	1	1	16
17	23	6	3	1	17
18	24	11	4	3	18
19	36	14	6	4	19
20	38	16	9	4	20
21	44	19	12	4	21
22	49	25	13	6	22
23	56	28	19	7	23
24	58	33	20	9	24
25	63	37	21	10	25
26	67	44	23	12	26
27	69	48	25	13	27
28	70	51	27	14	28
29	72	54	31	17	29
30	75	57	33	18	30
31	80	61	36	20	31
32	81	67	40	22	32
33	86	71	43	24	33
34	88	75	49	29	34
35	90	82	52	33	35
36	92	86	55	36	36
37	95	91	63	38	37
38	96	95	67	44	38
39	99	96	75	50	39
40	99	97	79	59	40
41	99	98	87	68	41
42	99	99	93	80	42
43	100	100	97	89	43
44	100	100	100	94	44
45	100	100	100	99	45
46	100	100	100	100	46

Raw Score	Percentiles for CGA-B				Raw Score
	P1	P2	P3	P4-P6	
0	1	1	1	1	0
1	1	1	1	1	1
2	1	1	1	1	2
3	1	1	1	1	3
4	1	1	1	1	4
5	1	1	1	1	5
6	1	1	1	1	6
7	2	1	1	1	7
8	3	1	1	1	8
9	4	1	1	1	9
10	5	1	1	1	10
11	6	2	1	1	11
12	7	2	1	1	12
13	11	3	1	1	13
14	13	4	1	1	14
15	15	4	2	1	15
16	20	5	3	1	16
17	26	7	5	2	17
18	36	9	6	2	18
19	39	14	7	3	19
20	45	18	11	4	20
21	49	22	13	5	21
22	51	26	16	8	22
23	56	29	17	9	23
24	57	35	18	10	24
25	60	41	19	12	25
26	65	42	22	13	26
27	67	51	26	15	27
28	74	55	28	16	28
29	74	60	31	18	29
30	75	61	34	20	30
31	81	65	35	22	31
32	83	70	39	26	32
33	86	75	44	29	33
34	89	78	51	30	34
35	90	79	55	32	35
36	91	87	61	37	36
37	92	89	65	41	37
38	95	92	70	45	38
39	98	95	75	49	39
40	99	95	82	58	40
41	100	99	88	68	41
42	100	100	94	80	42
43	100	100	97	88	43
44	100	100	98	97	44
45	100	100	100	99	45
46	100	100	100	100	46

9.2.2 An Initial Cut-off Point for Below Average Performance

With reference to the Clinical Evaluation of Language Fundamentals-Fifth Edition (CELF-5; Wiig, Semel, & Secord, 2013), a percentile rank of 16 or below, which is equivalent to “one standard deviation or below” in a normal distribution, is classified as a “below average” performance. According to this classification as described in Table 60, a percentile rank between 17th and 83the percentile ranks was considered “average” performance and a percentile rank of 84 or above was considered “above average” performance. The definition for “below average” performance is useful in identifying students with immediate needs of support for their development of grammatical knowledge in written Chinese. In the following section, we will test out if the definition is helpful in discriminating students with different abilities though statistical analysis.

Table 60. Level of performance according to projected equivalent percentile ranks of CGA raw scores

Level of performance	Percentile Ranks
Above Average Performance	≥ 84
Average Performance	17-83
Below Average Performance	≤ 16

9.2.2.1 Performance Levels of Students and CGA Raw Scores

To further verify if the set cutting points and the three CGA performance levels can help differentiate the students with different levels of abilities, one-way ANOVA using the Welch’s *F*-test was conducted. Post-hoc tests using Games-Howell procedure were also performed to check for the significance of mean differences of raw scores between the three CGA levels.

Results showed that there was significant main effect of CGA performance levels on CGA raw scores, with Welch's $F(2,311.45)=999.93$, $p<.01$, $\omega^2=0.539$ for CGA-A, and Welch's $F(2, 312.21)=1263.44$, $p<.01$, $\omega^2=.506$ (see Table 61). Post-hoc Games-Howell procedures revealed that there were significant mean differences between students with different levels of CGA performance, with a significant level of $p<.01$ (see Table 62). This implies that the abovementioned cutting points in Table 60 were effective in distinguishing students with different levels of performance in comprehending different grammatical knowledge in written Chinese.

Table 61. Results of Welch's *F*-test between CGA raw scores and performance levels (*N*=831)

	N	**Mean (*SD*)**	**Welch's *F***	***df*1**	***df*2**	**sig.**	**ω^2**
CGA-A	831	32.79 (8.76)	999.93**	2	311.45	.000	.539
CGA-B	831	32.51 (8.85)	1263.44**	2	312.21	.000	.506

**p<.05; **p<.01*

Table 62. Results of post-hoc Games-Howell pairwise comparisons between students' raw scores and different performance levels of CGA (*N*=831)

				Mean Difference	
CGA Levels	***N***	**Mean**	***SD***	**Below Average**	**Average**
CGA-A					
Below Average	117	19.18	4.851		
Average	548	33.05	6.751	-13.82**	
Above Average	166	41.53	3.150	-22.35**	-8.48**
CGA-B					
Below Average	121	19.57	4.474		
Average	586	33.05	7.022	-13.48**	
Above Average	124	42.57	2.586	-23.00**	-9.52**

**p<.05; **p<.01*

The above definition for students' CGA performance will be further reviewed based on the new set of DHH data. The focus is to examine if the cut-off points in percentile rank can help us identify students with (or without) significant difficulties in their development of grammatical knowledge in Chinese. Comparisons between students' percentile ranks and their background information and other performance data like academic scores among the three performance groups will also help us explore factors that may affect students' development in Chinese grammatical knowledge. A more detailed analysis will be reported in the following chapter.

Chapter 10: A Case Study with a Group of DHH Students

10.1 Reviewing the Performance of DHH Students Based on the Norms

With the different validity and reliability measures conducted, the two short tests of CGA, namely CGA-A and CGA-B, were confirmed with good validity and reliability for the assessment of Chinese grammatical knowledge in written Chinese. Therefore, the raw scores of the 27 DHH students assessed by CGA-A and CGA-B were converted to percentile ranks according to the two norms displayed in Table 59. As discussed in Section 9.2.2, the guidelines established in a language assessment, namely the Clinical Evaluation of Language Fundamentals-Fifth Edition (CELF-5; Wiig, Semel, & Secord, 2013) were adopted for the two CGA short tests. A percentile rank of 16th or below is classified as a "below average" performance. According to this classification, a percentile rank between 17th and 83th was considered "average" performance and a percentile rank ≥84th was considered as "above average" performance.

In this section, a case study would be held to test out whether this cut-off point is helpful in identifying DHH students with a relatively delayed development or below average performance in CGA, with reference to the other assessment results including the students' reading and writing scores, academic performance, and the results of LAMK.

As the reading and writing scores were based on different levels of Chinese Language examination at different grade levels, their raw scores were not comparable among each other. In this regard, the percentile ranks of each student in his or her class were calculated based on their raw scores (see results in Appendix D and Appendix E for the percentile ranks calculated for students' reading and writing scores by grade levels). The purpose is to find out the percentile ranks of the DHH students in compared with the performance of their classmates in his or her class who received the same extent of education and were taught by the same group of teachers.

10.1.1 Procedures

Following the above classification, the DHH students were grouped according to the percentile ranks of the students according to their results of the CGA short tests in raw scores. For those who got a percentile rank ≤ 16 for at least 3 out of 4 CGA test scores, they were classified as the "below average" group. DHH students who got at least 3 out of 4 CGA test scores ≥84 were classified as "above average". For the other students who did not fall into the above two groups, they were classified as students with "average" performance.

After classifying the DHH students into three groups with different levels of CGA abilities based on their percentile ranks, their background information and performance in other assessments would be compared to see if there were any specific differences among these three groups of DHH students. The observation may provide important information for us to explore possible factors that may affect the development of Chinese grammatical knowledge of DHH students.

10.1.2 Results

10.1.2.1 Students' Background and their CGA Performance

Table 63 summarizes the background information of the three groups of students based on the percentile ranks of their CGA scores. Among the 27 DHH students, 5 students who had CGA test scores ≤ 16 were grouped under "Below Average", and 6 students with the percentile rank ≥ 84 were grouped under "Above Average". The remaining 16 students were thus defined as having "Average" performance.

Table 63. Background of DHH students grouped by their levels of performance in CGA

		CGA Performance^ (N=27)							
Background Information		**Below Average (Percentiles ≤16) (N=5)**		**Average (Percentiles from 17-83) (N=16)**		**Above Average (Percentiles ≥84) (N=6)**		**Total (N=27)**	
		N	%	N	%	N	%	N	%
Gender	Male	3	20%	8	53%	4	27%	15	100%
	Female	2	17%	8	67%	2	17%	12	100%
Grade Levels	P1	0	0%	3	60%	2	40%	5	100%
	P2	0	0%	2	50%	2	50%	4	100%
	P3	1	17%	4	67%	1	17%	6	100%
	P4	1	33%	2	67%	0	0%	3	100%
	P5	1	25%	3	75%	0	0%	4	100%
	P6	2	40%	2	40%	1	20%	5	100%
Degree of Hearing Loss	Mild	0	0%	1	100%	0	0%	1	100%
	Moderate	0	0%	1	50%	1	50%	2	100%
	Moderately-severe	0	0%	2	67%	1	33%	3	100%
	Severe	0	0%	3	75%	1	25%	4	100%
	Profound	5	29%	9	53%	3	18%	17	100%
Other Disability#	Yes	2	100%	0	0%	0	0%	2	100%
	No	3	12%	16	64%	6	24%	25	100%
Hearing Device*	Nil	0	0%	1	100%	0	0%	1	100%
	HA	2	22%	4	44%	3	33%	9	100%
	CI (Bilateral)	2	22%	5	56%	2	22%	9	100%
	CI (Unilateral)	0	0%	4	80%	1	20%	5	100%
	ABI	1	33%	2	67%	0	0%	3	100%
Hearing Status of Parents	Deaf	3	75%	1	25%	0	0%	4	100%
	Hearing	2	9%	15	65%	6	26%	23	100%

^ CGA Performance is categorized based on percentiles projected by the raw scores of 831 TD data.

# DHH students' other disabilities such as Attention Deficit/Hyperactivity Disorder (ADHD) were confirmed after professional assessments.

** HA=hearing aids, CI=cochlear implants, and ABI=auditory brainstem implants.*

According to the background information of the 27 DHH students listed in Table 63, some major differences were observed between the three groups of students as described below. Since the sample size is small, and they were studying in a special education setting with both sign language and spoken language as the medium of instructions. The analysis below can only be a reference. No generalization of the observations can be made to other groups of DHH students in Hong Kong.

i) **Gender:** No gender difference on their CGA performance.

ii) **Grade Levels:** Most of the students had "Average" performance in CGA (from 40% to 75%). Relatively, there were more DHH students at the senior grade levels having "Below Average" performance (increased from 17% of P3 students to 40% of P6 students), and more students at the junior grade levels having "Above Average" performance (40% of P1 and 50% of P2 students, but 0-20% of P4-P6 students).

iii) **Degree of Hearing Loss:** No DHH students had "Below Average" performance in CGA, except those had profound hearing loss. Among the profound group, 29% (*N*=5) had "Below Average" performance, 53% (*N*=9) had "Average" performance, and 18% (*N*=3) had "Above Average" performance. The percentage in the "Below Average" was only slightly more than the "Above Average" group. As a whole, at least 50% of each group had an "Average" performance.

iv) **Other Disability:** 88% (*N*=22) of the DHH students with no additional disabilities had "Average" or Above Average" performance. However, all DHH students (100%, *N*=2) who had additional disability were in the "Below Average" group those the number is small.

v) **Hearing Device:** For all three groups of students, the DHH students were using different hearing devices including hearing aids, cochlear implants (unilateral or bilateral) and auditory brainstem implants. No specific difference could be found for the DHH students who were using hearing aids and bilateral cochlear implants. All DHH students who used unilateral cochlear implants had an “Average” or “Above Average” performance. For the three students using auditory brainstem implants, one had “Below Average” performance. The distribution was not distinctive among the whole group of DHH students.

vi) **Hearing Status of Parents:** Regarding parent status of the DHH students, 3 out of 4 students (75%) belonged to the “Below Average” group. The remaining one had “Average” performance. In contrast, only 2 out of 23 (9%) of the DHH students born to hearing parents had “Below Average” performance.

In sum, there is no gender difference between DHH students’ CGA performance. According to the results in Section 7.3.2 and Section 8.3.3, grade level has a significant main effect on CGA scores, which means that TD students at higher-grade levels have better CGA performance. However, even DHH students in this case study have enhanced performance at the higher-grade levels, their growth rates, as a whole, may not be comparable to TD students due to their hearing disability. Especially for those with profound hearing loss, they may face bigger challenge in their development of grammatical knowledge in written Chinese.

As mentioned in the background of the case study, all the DHH students in the case study are studying in a primary adopted the Sign Bilingualism and Co-enrollment in Deaf Education Programme. With the provision of sign language and spoken language as the medium of instructions in class, it is suggested that the barriers to communication in class can be reduced (Yiu, Tang, & Ho, 2019). According to Table 63, 71% of them had average or above average performance in CGA. Only 29% of them show greater difficulty in acquiring Chinese grammatical knowledge. However, for those with additional disabilities, their Chinese grammatical knowledge is relatively delayed. Additional support is necessary.

The choice of hearing device may not be a prominent factor affecting DHH students' performance in this case study. The hearing status of parents may be a factor that needs to be considered. Deaf parents in Hong Kong have long been depriving from accessible educational opportunity. In general, their educational background and socioeconomic status were relatively inferior when compared to hearing parents of DHH students. The academic or literacy support that deaf parents can give their DHH child may not be comparable to the hearing parents. Individual support to this group of DHH students is clearly identified from their CGA performance.

10.2 Significance of CGA Norms in Identifying Students in Need

As reported in Section 8.3.3.2, DHH students' CGA performance can significantly predict their normative academic performance in Chinese Language. As discussed in Section 9.2.2., the students' CGA performance were categorized as "Below Average" (percentile ranks ≤16), "Average" (percentile ranks 17-83) and "Above Average" (percentile ranks ≥84) to help identify students in need of support. In this section, more investigations will be focused on the effectiveness of the categorization in identifying students in need of academic support in Chinese Language.

The students' CGA scores as well as academic scores were thus put together, to see if students identified as "Below Average" in CGA would also show difficulties in their academic development in Chinese Language, including their reading and writing skills. Results in Table

64 reflects how DHH students' performance of CGA related to their academic performance in Chinese Language. The assumption is that students with "Below Average" performance in their Chinese grammatical knowledge are more likely to have delayed performance in their Chinese Language school examinations.

Table 64. Descriptive statistics that show the relationships between CGA's performance levels and the different academic scores

		CGA Performance^ (*N* =27)					
Academic performance		**Below Average (Percentiles ≤16) (*N* =5)**		**Average (Percentiles from 17- (*N* =16)**		**Above Average (Percentiles ≥84) (*N* =6)**	
		N	**%**	***N***	**%**	***N***	**%**
Academic Level (by LAMK)	Grade Appropriate	0	0%	13	81%	6	100%
	Below 1 year	0	0%	3	19%	0	0%
	Below 2 years	2	40%	0	0%	0	0%
	Below >2 years	3	60%	0	0%	0	0%
Reading (in percentiles#)	Above Average	0	0%	2	13%	2	33%
	Average	0	0%	12	75%	4	67%
	Below Average	5	100%	2	13%	0	0%
Writing (in percentiles#)	Above Average	0	0%	2	13%	4	67%
	Average	1	20%	12	75%	2	33%
	Below Average	4	80%	2	13%	0	0%

^ CGA Performance is categorized based on percentiles projected by the raw scores of 831 TD data
# The categories were defined as: "Below Average" =Percentiles ≤16, "Average"=Percentiles 17-83 and "Above Average"=Percentiles ≥84.

According to the results showed in Table 64, all the 5 (100%) students classified as "Below Average" in their CGA performance were showed to have at least two years delay in their academic attainment in Chinese Language by LAMK. For example, a P4 DHH student could only achieve a P2 or below standard in Chinese Language. For those categorized under "Average" performance, most of them achieved a grade-appropriate standard (*N*=13, 81%), only a few of them achieved a one-year-below standard (*N*=3, 19%). Moreover, all DHH students in the "Above Average" group (*N*=6, 100%) achieved their grade-appropriate standard.

Regarding the implications of DHH students' performance in CGA to their reading and writing performance in school examinations, results in Table 64 helps to get a clearer picture about their relationships. Firstly, all 5 DHH students under the "Below Average" category in CGA had "Below Average" performance in their reading comprehension (*N*=5, 100%), and 4 of them (80%) had "Below Average" performance in Chinese writing. Only one DHH student had average Chinese writing performance.

For DHH students had "Average" performance in CGA, 75% (*N*=12) of them had "Average" performance in their reading and writing performance in their Chinese Language school examination. Two of them (13%) were classified as "Below Average" and two of them in "Above Average" group in both Chinese writing and reading. For the "Above Average" group in CGA, their reading and writing scores were either in the "Average" (67%, *N*=4 in reading and 33%, *N*=2 in writing) or "Above Average" (33%, *N*=2 in reading and 67%, *N*=4 in writing) categories. No one DHH student was in the "Below Average" category.

With the above observation, the norms of CGA-A and CGA-B help to define the standard of performance in their Chinese grammatical knowledge with reference to a sample of 831 typically developing students in local primary schools. By using the raw scores of the two CGA short tests, educational or clinical practitioners can review DHH students' percentile ranks with reference to the performance of their same grade typically developing peers. This helps to get a clearer picture of how well the students perform in terms of their level of performance, especially those with below average performance who may need immediate support for their Chinese grammatical development. The major purpose of defining different levels of CGA performance into three categories is to help identify students who are significantly delayed in their CGA performance as soon as possible.

The significance of the cutting point of ≤16 percentile rank is essential for discriminating students in different abilities. According to the case study conducted in this study, DHH students identified as "Below Average" do show significant delayed development in their academic development, both in normative assessment by LAMK and school assessment in Chinese Language examination. This cutting point seems to be a good reference for practitioners to detect which students require immediate support and follow-up interventions. As a screening test, CGA-A and CGA-B shows to be effective in predicting the academic performance of both TD and DHH students in Chinese reading and writing. Further investigation with more DHH subjects would be helpful to further confirm the reliability of the categorizations.

Chapter 11: Conclusion

11.1 A Brief Review of the Study

The aim of the study is to develop a valid and reliable assessment tool for measuring Cantonese-speaking deaf and hard-of-hearing (DHH) students' grammatical knowledge in written Chinese. The data collected from the research project "Profiling Chinese Grammatical Knowledge of Deaf and Hard-of-Hearing Students in HK and China - A Comparative Study" (Tang et al., 2020) were used for the development of two short tests, functioning as screening tests for primary school students in Hong Kong. In order to develop a valid and reliable assessment tool for Chinese grammatical knowledge that is suitable for local Cantonese-speaking DHH students, a series of validation procedures were conducted (see Figure 7 for a summary of the procedures we have gone through in the validation and development process).

Ninety-two items from 18 grammatical categories and 46 sub-categories were selected from the original 172-item Chinese Grammatical Assessment (CGA) profiling tool after a series of psychometric reviews. Two equivalent lists of items were then selected for the establishment of two alternate forms of CGA, each comprised of 46 items. To ensure the two CGA short forms are reliable and valid, there were different psychometric evaluation through Rasch analysis, and a series of reliability and validity measures conducted for the initial dataset collected from 2015-2019, and also the newly collected dataset in 2022, including the tests for separation reliability, internal consistency, alternate forms reliability, and test-retest reliability, known-group validity, convergent validity, and construct validity for the validation of the two short forms of CGA, namely CGA-A and CGA-B. To avoid biased items toward either DHH or TD subjects, analysis for differential item functioning (DIF) were also conducted to collect more evidence before item selection for the two alternate forms.

In view that the two short tests would be used in different educational and clinical settings, two sets of norms in percentile ranks were established based on the 831 raw data of typically developing students from nine local primary schools in Hong Kong. With reference to the Clinical Evaluation of Language Fundamentals-Fifth Edition (CELF-5; Wiig, Semel, & Secord, 2013), three CGA performance levels were defined to categorize students' abilities in comprehending Chinese grammatical constructions. This categorization, especially the cutting point for "Below Average" performance, is helpful in identifying DHH students who are in need of immediate support or interventions. The result and its interpretations are more on educational or rehabilitation purposes, rather than a purely clinical diagnosis. Further studies should be conducted for the further verification of the cutting points adopted at this stage.

Prior validation procedures conducted for the two short tests confirmed that it is reliable to use the two alternate forms inter-changeably for comparisons. It is also feasible to use the two tests as repeated measures to track the students' development. Indeed, with the established predictive ability of CGA for both TD and DHH students' academic performance in Chinese, the assessment result can also be used as a reference for consideration of academic support to students in Chinese as a major subject in local primary schools.

11.2 A Summary of Research Findings

With the motivation of developing a normative assessment tool to measure Cantonese-speaking DHH students' grammatical knowledge in written Chinese, some research questions are set for the research:

i) Is the Chinese Grammatical Assessment (CGA) valid and reliable for measuring Chinese grammatical knowledge of Cantonese-speaking primary school children in

Hong Kong?

ii) Are the two CGA short tests comparable and reliable for assessing TD and DHH students’ grammatical knowledge in written Chinese?

iii) Are the norms set up for CGA effective in identifying DHH students who are in need of immediate support for Chinese grammatical development?

iv) Can CGA results be a significant predictor of DHH students’ academic performance in Chinese as a major subject in primary school?

In the following sections, we will summarize and discuss the major findings and their implications in this study, in response to the above research questions. Before further discussion made in the following sections, it is worth to note that since there are restricted availability of related literatures regarding the grammatical development of Cantonese-speaking DHH students in written Chinese, the depth and scope of discussions in this particular domain are relatively limited. In the following sections, the discussions will be focused on the different research questions raised for the study. The respective limitations of this study though pose difficulties for making more concrete conclusions and generalization of the findings. It indeed brings up important hypothesis, which should be addressed accordingly in future studies.

11.2.1 Is CGA Valid and Reliable?

Establishing test reliability and validity is essential in any test development. To achieve this, different areas of psychometric review are required. The application of Rasch analysis helps to evaluate and identify items that fit well with the model. Among the 172 items of the original design of CGA, 92 items are selected for the development of two 46-item short tests, and then a series of assessments were conducted to the test the reliability and validity of them. Reliability is concerned with the extent to which a measurement is consistent, and the test results are reproducible in various situations (Ng, 2014). Another important quality that an assessment or measurement should be established is about the assessment's validity. Whether the assessment is testing for what it intends to is the major concern of validity. Below is a brief summary of the results of different reliability conducted for CGA:

a) **Item/Person Separation Reliability:** The two CGA lists' person separation and reliability index got from Rasch analysis has confirmed that the items of CGA are able to distinguish the persons with different levels of ability reliably (see Section 7.2.1). In addition, the item separation and reliability measures of the two lists also prove that the items are having a good separation or distance between the items' difficulty levels, which in fact, also help estimate more accurately the persons' abilities (Wright & Stone, 1999). The results for person/item separation reliability were positive, with both item and person reliability > .80 and their separation index >2 for both alternate forms of CGA. Similar results were also found according to the dataset collected from DHH students.

b) **Internal Consistency:** The internal consistency measures based on Cronbach's alpha revealed that both 46-item CGA-A and CGA-B are having high internal consistency with the $\alpha > .90$ based on data of TD and DHH students (see Section 7.2.2).

c) **Test-retest Reliability:** Test-retest reliability was conducted for the test results of a smaller sample size with 102 TD subjects and 27 DHH subjects. Intraclass Correlation Coefficients (ICC) indicate that both CGA-A and CGA-B are having a "moderate to good" test-retest reliability (see Section 8.3.2.2).

Validity is not a clear-cut concept. There are many different ways to collect validity-related evidence, for example, good reliability is a necessary condition of test validity. In this study, different validity measures were conducted. The followings are the brief descriptions and results:

i) **Content Validity:** As reported in the above section, content validity of CGA was reviewed comprehensively by a group of 10 subject experts. The results are very positive with good endorsement from all the experts on the two CGA equivalent lists' representativeness, relevance and appropriateness of the assessment contents and items (see Section 11.2.2 for more detailed discussions).

ii) **Known-group Validity:** One-way analysis of variance was conducted to test the effect of grade level on CGA performance. The assumption is that students at higher grade levels will have better CGA test results because Chinese grammatical knowledge should have included as a part of the curriculum of Chinese Language in Hong Kong throughout the six-year primary education. Results indicated that there is a significant main effect of grade level on CGA performance (p<.01) (see Section 8.3.3.1). Post-hoc tests also found significant mean differences between all adjacent grade levels (p<.01) when the senior grade levels, i.e. P4 to P6, are grouped together as one single grade level, representing performance of students at senior primary level. According to the results, the known-group validity is established.

iii) **Convergent Validity:** Convergent validity of CGA is concerned with the relationships between CGA test scores and the other measure of related construct. In this study, though there is no other gold-standard measure used to test the TD students, their academic performance in Chinese Language as a major subject in primary education is considered a related construct to Chinese grammatical knowledge. The evaluation was conducted with a specific group of 102 TD and 27 DHH subjects. Statistical analyses confirmed that CGA performance is highly correlated with students' examination scores, in terms of reading comprehension and writing in Chinese. Significant correlation was found between all four CGA measures with the students' reading and writing scores (see Section 8.3.3.2). The correlation coefficient *r* ranged from .325 to .485, with a significant level of p<.01 for all measures of reading comprehension, and the coefficient ranged from .241 to .405, with a significant level of p<.01 for all measures of Chinese writing, indicating a good correlational relationship between CGA and their academic performance in Chinese Language.

For DHH subjects, standard scores of the Learning Achievement Measuring Kit (LAMK; Education Bureau, 2008, 2014) are also available for the test of convergent validity. Results indicate a significant correlational and predictive relationship between CGA scores and the normative academic scores in LAMK, with the correlation coefficients ranging from .783 to .838 (p<.01 for all measures). LAMK is a well-known academic assessment, providing both standard scores and equivalent grade-levels to determine the academic status of primary school students (Education Bureau, 2008, 2014). The results provide good evidence to support convergent validity of CGA for the assessment of both TD and DHH students.

iv) **Construct Validity:** Construct validity measures aim to evaluate how well a test measures what it is intended to measure. There is no one single evidence that can represent the overall construct validity of a measurement. Rather, it requires a wide range of consolidated evidence to confirm its significance. Rasch analysis has been conducted for the two CGA lists. Persons and items with outfit MNSQ out of the set range from 0.5-1.5 were deleted. The remaining items all fit well with the construct. As suggested by Aziz et al. (2014), the distributions of person ability and item difficulty displayed in the Wright map provides important information to support the tests' validity. Wright maps of both TD and DHH data help confirm the validity of the construct for CGA, only that more difficult item are required to match with the person ability of primary school students.

The analysis of Differential Item Functioning (DIF) of CGA is to avoid including items that are biased to either TD or DHH students. According to Boone, Staver and Yale (2014), this analysis also provides evidence to support the measurement's construct validity. After DIF analysis, some CGA items from the 172-item profiling tool were excluded from the two alternate forms after consideration also their results in other measures.

Besides results from Rasch analysis, the review of Known-group Validity and Convergent Validity for CGA also provides supporting evidence for CGA's Construct Validity. The measure for CGA's known-group validity proves that CGA is able to distinguish performance of students with different grade levels. On the other hand, the correlational and regression analysis between CGA scores and students' academic scores of both TD and DHH students provide positive evidence of convergent validity of the two CGA short tests.

11.2.1.1 Limitations and Future Developments

When a measurement is developed, the evaluation of its reliability and validity is essential. The two short versions of Chinese Grammatical Assessment (CGA) are developed based on a database with a set of collected data. Assessment data involving around one thousand subjects are collected from the study. The item pool with 172 items, following 18 grammatical categories, were designed for the assessment. The original intention of the 172-item profiling tool is to provide a summative assessment to explore how DHH students acquire different grammatical knowledge in written Chinese.

The current study is an extension of the original research objectives, trying to develop two normative short tests for educational and clinical purposes. Different reliability and validity measures, as mentioned above, are conducted. However, there are some limitations and constraints when the review was conducted. First of all, no other standardized assessments and database are available for the test of CGA's convergent validity except the data for their academic performance. Data for students' Cantonese grammar and reading comprehension with a "gold standard" would be essential to establish a stronger convergent validity of CGA. Though Cantonese grammar is different from that in written Chinese, but there should be a close relationship between students' oral language development and their literacy development. Especially for DHH students, with reference to Cummins (1989), students' acquisition of Cantonese as the first language should be able to provide them solid language concepts that are supportive to their development of grammatical knowledge in written Chinese as a second language.

According to the results of Rasch analysis, the distribution of item difficulty of the current

CGA items do not match with that of the person ability. More items with higher difficulty levels are required to help distinguish students with higher grade level or those with higher ability in Chinese grammatical knowledge. The whole construct will then be better established for both TD and DHH students with different levels of ability. The projected norms will also be able to precisely reflect the standards of the students participating in the assessment.

The DHH subjects involved in the study are all under the Sign Bilingualism and Co-enrollment in Education (SLCO) Programme. With the support of sign language, their overall language development may be different from other DHH students who do not have any input from sign language. Though the norms set for CGA are based on typically development students, the participation of DHH students from a wider education or developmental background can contribute a more representative sample for the establishment of test reliability and validity for CGA. Subjects like DHH students from different mainstream schools or special school for the deaf can be recruited for further confirmation of the effectiveness of the two CGA short tests.

11.2.2 Is CGA Representative, Relevant and Appropriate?

Whether CGA is representative, relevant and appropriate in its design for the intended purpose of the assessment requires comprehensive content validation. The original 172 items designed for CGA were reviewed before they were used for the initial data collection from students at different primary schools in Hong Kong. Three experts in Chinese linguistics and language acquisition had conducted the first phase of expert review on all items.

In view that the assessment would be turned into two short tests for educational and clinical use, 10 Subject Matters Experts (SMEs) including 5 speech therapists, and 5 teachers who have experience in teaching DHH students and Chinese were invited to help further review the content validity of the Chinese Grammatical Assessment from their perspectives (see Chapter 4). The main purpose of this expert panel is to review the representativeness, relevance and appropriateness of the grammatical categories involved and the respective test items designed for CGA. The review also includes the administration and operations of CGA.

The results are reflected by the SMEs' ratings and the projected content validity index (CVI). CVI represents the proportion of SMEs endorse the content of the assessment with a range of scores from 0-1.0. CVI=1.0 represents a perfect score with all SMEs endorsing the items or the operational elements of CGA. In sum, the ratings given by the 10 SMEs are very positive, which means that they highly endorse the representativeness of the selected 18 grammatical categories in Chinese (CVI=.90) for item development and the average CVIs for the relevance and appropriateness of the 172 items in CGA are .91 and .95 respectively. When the two alternate forms of CGA short tests were confirmed, the average CVIs regarding relevance and appropriateness of the items were reviewed with a result of .96 and .92 for CGA-A, and .91 and .95 for CGA-B respectively.

11.2.2.1 Limitations and Future Developments

The 10 Subject Matter Experts (SMEs) have not only provided review ratings for different content areas of CGA, but also given useful comments and suggestions for the assessment content and item design. Moreover, there are also lots of practical suggestions and considerations for the assessment's overall administration and operations, even the name of the assessment. Not all suggestions have been adopted before data collection for the new dataset.

After the review, some items are excluded during the item selection process for the two short tests in consideration of the results of other measures. According to the SMEs' comments, some further development of reliable and valid items are required. The item design such as the pictures of the stimuli, the structure of the sentences or the semantic implications of the stimuli can be further modified for enhancement of CGA's content validity. Further discussions are required to review all the suggestions and consolidate a list of agreed tasks with priority.

11.2.3 Are the Two CGA Tests Comparable and Reliable?

In view that two equivalent lists of CGA have been developed for alternating use in different situations, for example, a close follow up of students' progress or a confirmation of results through repeated testing. The Alternate Forms Reliability was evaluated based on both norming data and the new sets of data collected from a group of TD and DHH students. Results of Intraclass Correlation Coefficients (ICC) indicate that CGA-A and CGA-B are having "good to excellent" reliability and comparable results between the two short tests according to the standard recommended by Koo and Li (2016), and the positive results apply to both TD and DHH subjects (see Section 7.2.3). The results of CGA-A and CGA-B are comparable and highly correlated with other related performance such as academic performance in Chinese language.

11.2.3.1 Limitations and Future Developments

Currently, the alternate forms reliability of the two CGA short tests have basically been established with good results according to the first set of data collected from TD and DHH subjects using the 172-item version. The reliability coefficient are less positive in the separate data collection from the TD students participating in the Sign Bilingualism and Co-enrollment Programme (.684, with the 95% CI between .566 to .775 for the first round of assessment and 832, with the 95% CI interval between .760 and .833). It would be conducive to have more primary school students tested with the two short tests separately to further establish the alternate forms reliability of the two tests, especially when further modifications of the two CGA lists will be done.

11.2.4 Are the Norms of CGA Effective in Identifying DHH Students in Need?

The norms of both short tests of CGA have been set up based on the norming data of 831 typically developing students. As the assumption of normality of the data was violated, the raw scores were converted to percentile ranks for reference. Based on the norms of TD students, we can have a good reference to know how well an individual DHH student performs based on the percentile rank he or she is positioned at his or her grade level. With reference to the Clinical Evaluation of Language Fundamentals-Fifth Edition (CELF-5; Wiig, Semel, & Secord, 2013), a percentile rank of 16 or below, which is equivalent to “one standard deviation or below” in a normal distribution, is classified as a “below average” performance. Applying this definition to the 27 DHH students from a mainstream school, the five students who were classified as having “Below Average” CGA performance are all confirmed to have a at least two-year delay in a normative academic assessment by LAMK (Education Bureau, 2008, 2014). Moreover, the “Below Average” group also has a delayed academic development, reflected by their Chinese reading and writing scores in school examinations.

11.2.4.1 Limitations and Future Developments

The norms for the two CGA short tests are established with 831 TD data, but the distribution of the dataset from different grade levels is not balanced especially when students from P4 to P6 are grouped together to form a senior primary group. In addition, the norms of the assessment would be more robust in distinguishing students with different ability when more difficult items can be included in the item pools so that students' ability from P4-P6 can be distinguishable, and the norms for individual grade levels from P4 to P6 can be set up for a more accurate review of students' abilities. It can also help identifying possible factors that may affect their development of Chinese grammatical knowledge of DHH students such as degree of hearing loss, additional disabilities, or parents' hearing status.

As discussed before, the cutting point (≤16 percentile ranks) is set for identifying students with "Below Average" performance so that teachers and clinicians can support DHH students in need more effectively. The cutting point should be further reviewed when more data are collected for further analysis.

11.2.5 Can CGA Results be a Significant Predictor of Academic Performance?

In view that Chinese grammatical knowledge is associated positively with reading comprehension and writing scores of Chinese Language examination, linear regression was conducted to check if the two constructs are correlated with each other. In addition, the predictive power of CGA on students' academic performance in Chinese was also investigated. The analysis reveals that CGA scores can significantly predict the examination results in Chinese Language. The significant results further confirm the convergent validity of CGA and reiterate the significance of the development of CGA, which helps to review Cantonese-speaking TD and DHH students' grammatical knowledge in written Chinese.

11.2.5.1 Limitations and Future Developments

The significance in predicting academic performance in Chinese Language from CGA scores is helpful to further confirm the needs of immediate support to students with poor performance in CGA. The case study is only based on DHH students participating in the Sign Bilingualism and Co-enrollment in Education Programme. No generalization to other DHH students or other student populations can be made. The use of LAMK as a gold standard in academic assessment for both TD and DHH students with a wider educational background would be an effective way to further establish a more solid predictive relationship between CGA and academic attainment in Chinese Language. This would extend the educational and clinical applications and significance of the assessment in the long run.

11.3 Educational and Clinical Applications

11.3.1 Identifying Students in Need and the Mode of Interventions

Chinese Language is an important area of development that a student in Hong Kong has to achieve. No matter students who are typically developing or having some special needs, their management of the curriculum helps them accomplish different areas of learning as the test books are mostly written in Chinese. With the establishment of CGA, and its significance in predicting students' academic achievements in Chinese Language, the impact of Chinese grammatical knowledge as a foundation of Chinese literacy warrants a long-term impact to the academic development of students in Hong Kong.

In this study, the Chinese Grammatical Assessment has established to review students' grammatical knowledge in written Chinese, which predicts their academic development in Chinese Language. Difficulties in acquiring Chinese grammatical knowledge or a delayed development would affect the students' academic performance in schools. Many students with hearing loss are mainstreamed in regular schools failed to perform well in Chinese Language, however, the inventions may be academically-oriented like attending remedial tutorials and doing lots of exercise in order to raise their standards. According to the current study, developing grammatical knowledge in written Chinese should be included as one of the strategic support for this group of students. With the enhancement of their respective knowledge in Chinese grammar, their overall academic performance may also be enhanced.

11.3.2 Early Identification and Interventions

The original CGA profiling tool includes 172 items. Students are required to attend the assessment in a few sessions before they can complete all items attentively. After developing the two CGA short tests with only 46 items, 15-20 minutes are sufficient for a student to complete one test. The CGA short tests are effective screening tools that help identify students' needs in Chinese grammatical development. Early identification always comes with early interventions. With the implementation of the two CGA short tests as screening tools for delayed Chinese grammatical development, DHH students with the needs for supported can thus be identified early. Effective early intervention can thus be more guaranteed.

11.3.3 Developing Computer-aided Tests (CAT)

In order to provide a more efficient screening assessment for students, developing a computer-aided test (CAT) for CGA may be a valuable future development (Canon et al., 2020). The advantage of CAT is that it can provide test items with appropriate a difficulty level for a particular student based on a pool of validated items with defined difficulty levels. Computer programming can help assign items with a higher or lower difficulty level to a student based on his or her prior responses. The pre-registered difficulty levels of the items can be developed based on a set of normed data and any new data continuously collected from targeted candidates. Rasch's item-level analysis including item difficulty and fitness would be conducive to the development of CAT for CGA in future.

11.3.4 Outcome Measures

With the development of CGA, educators can include the two short tests as a screening tool to profile the learning or language outcomes of DHH students. In addition to academic outcomes, we can take a broader perspective to observe the students' development in Chinese grammatical knowledge. The availability of two short tests can provide teachers or clinicians with a flexible use of the two tests to track the progress of the students. The objective norms established can provide reliable reference for a more accurate interpretation of students' outcomes.

Grammatical competence is a significant factor affecting DHH students' reading ability (Kelly, 1996). From a language development perspective, the tests can also be a tool to observe students' literacy development. The item responses can also be a good reference for the speech and language therapists to understand which categories of grammatical knowledge a student may be more vulnerable to them.

CGA will possibly be a measure for the outcomes of a specific intervention or more broadly a deaf education programme like the Sign Bilingualism and Co-enrollment in Education Programme, no matter in the inclusive or special school settings. With some further research on CGA's applications in different special needs populations, it will be a useful tool with significant practical and clinical use for a wide spectrum of special needs.

11.4 Conclusion

Ineffective grammatical development in a language is a long-standing problem facing deaf and hard-of-hearing (DHH) children (Quigley et. al. 1976; Wilbur, Goodhart & Montandon, 1983; Berent 1988, 1996; de Villiers, de Villiers & Hoban, 1994; Lillo-Martin 1998; Friedmann & Szterman, 2006, 2011; Volpato, 2010; Guasti et al., 2014; Yiu, 2004, 2012; Lam, 2015, and among others). The impact is not only on DHH children's reading and literacy development (Kelly, 1996), it also affects their academic performance (Babbidge, 1965; Holt, 1993; Traxler, 2000; Qi & Mitchell, 2012). In Hong Kong, how Cantonese-speaking DHH children acquires grammatical knowledge in written Chinese, which follows a different grammatical system from Cantonese is inevitably a complex issue yet to be explored. Language deprivation (Lau et, al., 2019) and academic failure (The Hong Kong Society for the Deaf, 2009) have been the phenomena commonly observed in local deaf community in Hong Kong. The development of the Chinese Grammatical Assessment (CGA) is to support educational and clinical professionals to understand better the needs of DHH students in their reading or literacy development. The two normative alternate forms of CGA are established through a series of psychometric evaluations that help to collect empirical evidence for the establishment of two valid and reliable short tests. A standardization process helps to create the norms of the two short tests in percentile ranks an respective standards to discriminate students with different abilities.

There are different limitations the study comes across during the development process. Further item refinement and data collection are conducive to collecting more evidence for the tests' development. Specifically, more items with higher difficulty levels are required to distinguish students with higher person ability. Some different grammatical structures may also be included according to the suggestions from the SMEs to ensure a more comprehensive coverage of grammatical knowledge that is appropriate for the assessment of primary school students.

Further research may focus on developmental pathway for the different grammatical knowledge and the differences that may exist between the acquisition of grammatical knowledge in written Chinese in DHH and typically developing students. Other than the application of CGA in students with hearing loss, its applications in other different special needs students may also be a possible development worth to be explored in future.

Reference

Anselmi, P., Colledani, D., & Robusto, E. (2019). A comparison of classical and modern measures of internal consistency. *Frontiers in Psychology, 10,* 1-12. **https://doi.org/10.3389/fpsyg.2019.02714**

Aryadoust, V., Ng, L.Y., & Sayama, H. (2021). A comprehensive review of Rasch measurement in language assessment: Recommendations and guidelines for research. *Language Testing, 38*(1), 6-40.

Audit Commission (2018) Chapter 3: Education Bureau-Integrated Education. Hong Kong: Audit Commission.

Abdul Aziz, A., Jusoh, M.S., Omar, A.R., Amlus, M.H., & Awang Salleh, T.S. (2014). Construct validity: a Rasch measurement model approaches. *Journal of Applied Science and Agriculture, 9*(12), 7-12.

Ball, M. J., Crystal, D., & Fletcher, P. (Eds.). (2012). *Assessing Grammar: The Languages of LARSP*. Multilingual Matters.

Barr, R., Sadow, M., & Blachowicz, C. (2002). *Reading diagnosis for teachers: An instructional approach*. Allyn and Bacon.

Babbidge, H. D. (1965). *Education of the Deaf: A Report to the Secretary of Health, Education, and Welfare by His Advisory Committee on the Education of the Deaf.* Retrieved from, **https://files.eric.ed.gov/fulltext/ED014188.pdf**

Berent, G. P. (1988). An assessment of syntactic capabilities. In M. Strong (Eds.), *Language Learning and deafness* (pp. 133-61). Cambridge University Press.

Berent, G. P. (1996). Syntax Acquisition by Deaf Learners. In W. C. Ritchie, & T. K. Bhatia (Eds.) *Handbook of Second Language Acquisition* (pp. 469-506). Academic Press.

Berent, G. P. (2004). Sign language-spoken language bilingualism: code mixing and mode mixing by ASL-English. In T. K. Bhatia & W. C. Ritchie (Eds.), *The Handbook of Bilingualism* (pp. 312-335). Blackwell Pub.

Berent, G.P. & Kelly, R.R. (2008). The efficacy of visual input enhancement in teaching deaf learners of L2 English. In Z-H. Han (Eds.), *Understanding second language process* (pp. 80-105). Multilingual Matters.

Blamey, J.P., Sarant, Z.J., Paatsch,E.L., Barry,G.J., Bow,P.C., & Wales,J.R. (2001). Relationships among speech perception, production, language. *Journal of Speech*, *44*, 264-285.

Bond, T.G. & Fox, C.M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences* (2nd ed.). Lawrence Erlbaum Associates, Inc., Publishers.

Boone, W. J. (2016). Rasch Analysis for Instrument Development: Why, When, and How? *CBE Life Sciences Education, 15*(4), rm4. **https://doi.org/10.1187/cbe.16-04-0148**

Brisbois, J. E. (1995). Connections between first-and second-language reading. *Journal of Reading Behavior*, *27*(4), 565-584.

Bus, A. G., & van IJzendoorn, M. H. (1999). Phonological awareness and early reading: A meta-analysis of experimental training studies. Journal of Educational Psychology, *91*(3), 403–414. **https://doi.org/10.1037/0022-0663.91.3.403**

Cannon, J. E., Easterbrooks, S. R., Gagné, P., & Beal-Alvarez, J. (2011). Improving DHH students' grammar through an individualized software program. J*ournal of Deaf Studies and Deaf Education*, *16*, 437-457.

Cannon, J. E., & Hubley, A. M. (2014). Content validation of the Comprehension of Written Grammar assessment for deaf and hard of hearing students. *Journal of Psychoeducational Assessment*, *32*(8), 768-774.

Cannon, J. E., Hubley, A. M., Millhoff, C., & Mazlouman, S. (2016). Comprehension of written grammar test: reliability and known-groups validity study with hearing and deaf and hard-of-hearing students. *Journal of Deaf Studies and Deaf Education*, *21*(1), 54-63.

Cannon, J. E., Hubley, A. M., O'Loughlin, J. I., Phelan, L., Norman, N., & Finley, A. (2020). A technology-based intervention to increase reading comprehension of morpho-syntactic structures. *Journal of Deaf Studies and Deaf Education*, 25(1), 126-139. **https://doi.org/10.1093/deafed/enz029**

Castellanos I., Pisoni D.B., Kronenberger W.G., Beer J. (2016). Early Expressive Language Skills Predict Long-Term Neurocognitive Outcomes in Cochlear Implant Users: Evidence from the MacArthur-Bates Communicative Development Inventories. *American Journal of Speech-Language Pathology, 25*(3), 381-382.

Census and Statistics Department. (2022). *2021 Population Census*. Retrieved from: https://www.censtatd.gov.hk/en/EIndexbySubject.html?scode=600&pcode=D5212101

Chan, Y. C., & Yang, Y. J. (2018). Early reading development in Chinese-speaking children with hearing loss. *Journal of Deaf Studies and Deaf Education*, *23*(1), 50-61. **https://doi.org/10.1093/deafed/enx042**

Chen, H. C. (1996). Chinese reading and comprehension: A cognitive psychology perspective. In M. H. Bond (ed.), *The Handbook of Chinese Psychology* (pp. 43–62). Oxford University Press.

Chen, K., Li, B.Y., & Sun, L. (2016). On the acquisition of Hearing-impaired students' written tendency verbs. *Chinese Journal of Special Education*. *188*(2), 43-48.

Cheung, K.Y., Leung, M., & McPherson, B. (2013). Reading strategies of Chinese students with severe to profound hearing loss. *Journal of Deaf Studies and Deaf Education*, *18*(3), 312-28.

Ching, T. Y., Dillon, H., Marnane, V., Hou, S., Day, J., Seeto, M., Crowe, K., Thomas, J., Van Buynder, P., Zhang, V., Wong, A., Burns, L., Flynn, C., Cupples, L., Cowan, R.S.C., Sjahalam-King, J., & Yeh, A. (2013). Outcomes of early-and late-identified children at 3 years of age: findings from a prospective population-based study. *Ear and Hearing*, *34*(5), 535. **https://doi.org/10.1097/AUD.0b013e3182857718**

Ching, B. H. H., & Nunes, T. (2015). Concurrent correlates of Chinese word recognition in deaf and hard-of-hearing children. *Journal of Deaf Studies and Deaf Education*, *20*(2), 172-190.

Choi, I., & Papageorgiou, S. (2020). Evaluating Subscore Uses Across Multiple Levels: A Case of Reading and Listening Subscores for Young EFL Learners. *Language Testing, 37* (2), 254-279. **https://doi.org/10.1177/0265532219879654**

Chomsky, N. (1957). *Syntactic structures*. Mouton.

Clarke, E. R., Rogers, W. T., & Todd, W. (1981). Correlates of Syntactic Abilities in Hearing-Impaired Students. *Journal of Speech and Hearing Research*, 24, 48-54.

Chung, K.K.H., McBride-Chang, C., Wong, S.W.L., Cheung, H., Penney, T.B., & Ho, C. S.-H. (2008). The role of visual and auditory temporal processing for Chinese children with developmental dyslexia. *Annals of Dyslexia, 58*, 15-35. **https://doi.org/10.1007/s11881-008-0015-4**

Colletti, V. & Shannon, R. V. (2005). Open Set Speech Perception with Auditory Brainstem Implant? *Laryngoscope, 115,* 1974-1978.

Davidson, M. (2014). Known-groups validity. In A.C., Michalos (Ed.) *Encyclopedia of Quality of Life and Well-being Research* (pp. 3481-3482). Springer.

Dragounova, Z (2018). Development and standardization of a rating scale designed for floorball skills diagnostics of young school-age children. *Baltic Journal of Health and Physical Activity, 10*(4), 34-48. **https://doi.org/10.29359/BJHPA.10.4.03**

De Villiers, J., de Villiers, P., & Hoban, E. (1994). The Central Problem of Functional Categories in the English Syntax of Oral Deaf Children. In H. Tager-Flusberg (Ed.), *Constraints on Language Acquisition Studies of Atypical Children* (pp. 9-47). Lawrence Erlbaum Associates.

Duchesne, L. (2016). Grammatical competence after early Cochlear Implantation. In M. Marschark, & P. E. Spencer (Eds.), *The Oxford Handbook of Deaf Studies and Education*, (pp. 113-131). Oxford University Press.

Easterbrooks, S. R., Lederberg, A. R., Miller, E. M., Bergeron, J. P & Connor, C. M. (2008). Building the alphabetic principle in young children who are deaf or hard of hearing. *Volta Review*, *109*(2-3), 87-119.

Easterbrooks, S. R. (2010). *Comprehension of written grammar*. [Unpublished assessment] Department of Educational Psychology and Special Education, Georgia State University, Atlanta, Georgia.

Education Bureau (2008). *Learning Achievement Measurement Kit 2.0: User's Guide*. Education Bureau.

Education Bureau (2012). The EDB circular (EDBC012/2012) on learning support. Education Bureau.

Education Bureau (2014). *Learning Achievement Measurement Kit 3.0: User's Guide.* Education Bureau.

Engen, E., & Engen, T. (1983). *Rhode Island Test of Language Structure*. Pro-Ed, Incorporated.

Figueras, B., Edwards, L., & Langdon, D. (2008). Executive function and language in deaf children. *Journal of Deaf Studies and Deaf Education, 13*, 362–377. **https://doi.org/10.1093/deafed/enm067**

Flanagan, J.C. (1951). Units, scores, and norms. In E.F. Lindquist (Ed.) *Educational Measurement* (pp. 695-763). American Council on Education.

Flesch, R. (1955). *Why Johnny Can't Read*. Harper Collins.

Friedmann, N., & Szterman, R. (2006). Syntactic movement in orally-trained. *Journal of Deaf Studies and Deaf Education, 11*, 56-75.

Friedmann, N., Novogrodsky, R., Szeterman, R., & Preminger, O. (2008). Resumptive pronouns as a last resort when movement is impaired: Relative clauses in hearing impairment. In S. Armon-Lotem, S. Rothstein, & G. Danon (Eds.), *Generative Approaches* *to Hebrew Linguistics*. (pp.267-290). John Benjamins.

Friedmann, N., Szterman, R., & Haddad-Hanna, M. (2009). The Comprehension of Relative Clauses and Wh Questions in Hebrew and Palestinian Arabic Hearing Impairment. In J. Costa, M. L. Castro, & F. Pratas (Eds.), *Language Acquisition and Development: Generative Approaches to Language Acquisition* (pp. 2-12). Cambridge Scholars Press/CSP.

Friedmann, N., & Szterman, R. (2011). The Comprehension and Production of Wh-questions in Deaf and Hard-of-Hearing Children. *Journal of Deaf Studies and Deaf Education*, *16*(2), 212-235. **https://doi.org/10.1093/deafed/enq052**

Gaustad, M. G., & Kelly, R. R. (2004). The relationship between reading achievement and morphological word analysis in deaf and hearing students matched for reading level. *Journal of Deaf Studies and Deaf Education*, *9*, 269–285. **https://doi.org/10.1093/ deafed/enh030**

Geers, A. E. (2003). Predictors of reading skill development in children with early cochlear implantation. *Ear and Hearing*, *24* (1 Supp), 59S-68S. **https://doi.org/ 10.1097/01.AUD.0000051690.43989.5D**

Geers, A. (2004). The ears of the deaf unstopped. *Seminars in Hearing*, *25* (3), 257-268.

Geers, A., Moog, J. S., Biedenstein, J. J., Brenner, C., & Hayes, H. (2009). Spoken language scores of children using cochlear implants compared to hearing age-mates at school entry. *Journal of Deaf Studies and Deaf Education*, 14, 372–385. **https://doi.org/ 10.1093/deafed/enn046**

Gerken, L., & Shady, M. E. (1996). The picture selection task. In D. McDaniel, C. McKee, & H. S. Cairns (Eds.), *Methods for Assessing Children's Syntax* (pp. 125–145). The MIT Press.

Gioia, G.A., Espy, K.A., & Isquith, P.K. (2003). *Behavior Rating Inventory of Executive Function-Preschool Version: Professional Manual*. PAR.

Gordon, P. (1996). Truth-Value Judgment. D. McDaniel, C. McKee, H. Cairns (Eds.) *Methods for Assessing Children's Syntax* (pp. 211-242). The MIT Press.

Guasti, M. T. (2002). *Language Acquisition: The Growth of Grammar.* The MIT Press.

Guasti, M. T., Papagno, C., Vernice, M., Cecchetto, C., Giuliani, A., & Burdo, S. (2014). The effect of language structure on linguistic strengths and weaknesses in children with cochlear implants: Evidence from Italian. *Applied Psycholinguistics*, *5*(4), 739–764. **https://doi.org/10.1017/S0142716412000562**

Guilford J. P. (1965). *Fundamental Statistics in Psychology and Education* (4th ed.). McGraw-Hill.

Hall, M. L., Eigsti, I-M., Bortfeld, H., & Lillo-Martin, D. (2017). Auditory deprivation does not impair executive function, but language deprivation might: Evidence from a parent-report measure in deaf native signing children. *Journal of Deaf Studies and Deaf Education*, *22,* 9–21.

Hambleton, R., & Jones, R. (1993). Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practice, 12* (3), 38-47.

Hammond, S. (1995). Using psychometric tests. In G. M. Breakwell, S. Hammond, & C. Fife-Schaw (Eds.) *Research Methods in Psychology* (pp.194-212). Publications Ltd.

Harris, M., & Beech, J. R. (1998). Implicit phonological awareness and early reading development in prelingually deaf children. *The Journal of Deaf Studies and Deaf Education*, *3*(3), 205-216.

Harris, M., Terlektsi, E., & Kyle, F. E. (2017). Literacy out- comes for primary school children who are deaf and hard of hearing: A cohort comparison study. *Journal of Speech, Language, and Hearing Research*, 60, 701–711. **https://doi.org/10.1044/2016_JSLHR-H-15-0403**

Hay-McCutcheon, M. J., Kirk, K. I., Henning, S. C., Gao, S., & Qi, R. (2008). Using early language outcomes to predict later language ability in children with cochlear implants. *Audiology and Neurotology*, *13*, 370–378. **https://doi.org/10.1159/000148200**

Hermans, D., Knoors, H., Ormel, E., & Verhoeven, L. (2008). The relationship between the reading and signing skills of deaf children in bilingual education programs. *Journal of Deaf Studies and Deaf Education*, *13*(4), 518-530.

Hohensinn, C., & Kubinger, K. D. (2011). Applying item response theory methods to

examine the impact of different response formats. *Educational and Psychological Measurement, 71*(4), 732–746. **https://doi.org/10.1177/0013164410390032.**

Holmefur, M., Aarts, P.B., Hoare, B., & Krumlinde-Sundholm, L. (2009). Test-retest and alternate forms reliability of the Assisting Hand Assessment. *Journal of Rehabilitation Medicine, 41*(11), 886-91. **https://doi.org/10.2340/16501977-0448**

Holt, J. A. (1993). Stanford Achievement Test—8th edition: Reading comprehension subgroup results. *American Annals of the Deaf*, *138*(2), 172-175.

Hong Kong Government. (1977). *Integrating the Disabled into the Community: A United Effort.* Government Printer.

Hoover, W., & Gough, P. (1990). The Simple View of Reading. *Reading and Writing, 2*(2), 127-160.

Huang, C.-T. J., Li, Y.-H. A., & Li, Y. (2009). *The Syntax of Chinese*. Cambridge University Press.

Humphries, T., Kushalnagar, P., Mathur, G., Napoli, D. J., Padden, C., Rathmann, C., & Smith, S. R. (2012). Language acquisition for deaf children: Reducing the harms of zero tolerance to the use of alternative approaches. *Harm Reduction Journal*, *9*(1), 1-9.

Hyams, N. (1987). The theory of parameters and syntactic development. In T. Roeper & E. Williams (Eds.), *Parameter setting* (pp. 1-22). Reidel.

James, D., Rajput, K., Brinton, J., & Goswami, U. (2008). Phonological awareness, vocabulary, and word reading in children who use cochlear implants: Does age of implantation explain individual variability in performance outcomes and growth?. *Journal of Deaf Studies and Deaf Education*, *13*(1), 117-137.

Jamieson, J. R. (1994). The Impact of Hearing Impairment. In J. Katz (Ed.), *Handbook of Clinical Audiology* (4th ed.) (pp. 596-615). Williams & Wilkins.

Kelly, L. (1996). The interaction of syntactic competence and vocabulary during reading by deaf students. *The Journal of Deaf Studies and Deaf Education*, *1*(1), 75-90.

Ku, Y. M., & Anderson, R. C. (2003). Development of morphological awareness in Chinese and English. *Reading and Writing*, *16*(5), 399-422.

Lam, S. W. Z. (2015). Acquisition of Cantonese relative clauses by typically developing and deaf children. In J. T. Sun, & Y. M. Yao (Eds.), *Proceeding of the 18th Internal Conference on Yue Dialects* (pp. 244-277). Jinan University Publisher.

Lange, C. M., Lane-Outlaw, S., Lange, W. E., & Sherwood, D. L. (2013). American Sign Language/English bilingual model: A longitudinal study of academic growth. *Journal of deaf studies and deaf education*, *18*(4), 532-544.

Larsen-Freeman, D., & Celce-Murcia, M. (2016). *The Grammar Book: Form, Meaning, and Use for English Language Teachers.* (3rd ed.). National Geographic Learning, Heinle Cengage Learning.

Lau, T. H. M., Lee, K. Y. S., Lam, E. Y. C., Lam, J. H. S., Yiu, C. K. M., & Tang, G. W. L. (2019). Oral language performance of deaf and hard-of-hearing students in mainstream schools. *The Journal of Deaf Studies and Deaf Education*, *24*(4), 448-458.

Lebeaux, D. (1987). Comments on Hyams. In T. Roeper & E. Williams (Eds.), *Parameter setting* (pp. 23-30). Reidel.

Lederberg, A. R., Schick, B., & Spencer, P. E. (2013). Language and literacy development of deaf and hard-of-hearing children: Successes and challenges. *Developmental Psychology*, *49*(1), 15–30. **https://doi.org/10.1037/a0029558**

Lee, K. Y., van Hasselt, C. A., & Tong, M. C. (2010). Age sensitivity in the acquisition of lexical tone production: evidence from children with profound congenital hearing impairment after cochlear implantation. *Annals of Otology, Rhinology & Laryngology*, *119*(4), 258-265.

Lee, Y. M., Kim, L. S., Jeong, S. W., Kim, J. S., & Chung, S. H. (2010). Performance of children with mental retardation after cochlear implantation: speech perception, speech intelligibility, and language development. *Acta Oto-laryngologica*, *130*(8), 924-934.

Lee, W., Kim, S. Y., Choi, J., & Kang, Y. (2019). IRT approaches to modeling scores on mixed-format tests. *Journal of Educational Measurement*. **https://doi.org/10.1111/jedm.12248**

Li, C. N., & Thompson, S. A. (1981). *Mandarin Chinese: a functional reference grammar*. Berkeley: University of California Press.

Li, Q. (2015). *Acquisition of Chinese passives by deaf learners.* [Unpublished master's thesis], The Chinese University of Hong Kong.

Lillo-Martin, D. (1992). Deaf readers and Universal Grammar. In M. Marschark & D. Clark (Eds.), *Psychological Perspectives on Deafness* (pp. 311-337). Lawrence Erlbaum Associates.

Lillo-Martin, D. (1998). The acquisition of English by deaf signers: Is Universal Grammar involved? In S. Flynn, G. Martohardjono, & W. O'Neil (Eds.), *The Generative Study of Second Language Acquisition* (pp. 131-149). Lawrence Erlbaum.

Linacre, J. M. (1989). *Many-facet Rasch measurement*. MESA Press.

Linacre, J. M. (1995). Reliability and separation nomograms. *Rasch Measurement Transactions*, *9*(2), 421. **https://www.rasch.org/rmt/rmt92a.htm**

Linacre, J. M. (2002). What do Infit and Outfit, mean-square and standardized mean? *Rasch Measurement Transactions, 16*, 878.

Linacre, J. M. (2005). Rasch dichotomous model vs. one-parameter logistic model. *Rasch Measurement Transactions, 19*(3), 1032.

Linacre, J.M. (2012). *A user's guide to Winsteps. Ministeps. Rasch-model computer programs. Program manual 3.74.0.* https://www.winsteps.com/a/Winsteps-Manual.pdf

Lord, F. M. (1980). *Applications of Item Response Theory to Practical Testing Problems.* Lawrence Erlbaum Associates, Inc.

Lord, F. M., & Novick, M. R. (1968). Statistical theories of mental test scores. *Reading*, Addison-Wesley.

Lord, F. M., & Wright, B. D. (2010). Fred Lord and Ben Wright discuss Rasch and IRT models. *Rasch Measurement Transactions, 24*, 1289–1290.

López-Higes, R., Gallego, C., Martín-Aragoneses, M. R., & Melle, N. (2015). Morpho-Syntactic Reading Comprehension in Children with Early and Late Cochlear Implants. *Journal of Deaf Studies and Deaf Education, 20* (2), 136-146. **https://doi.org/10.1093/deafed/env004**

Lust, B. (2006). *Child Language: Acquisition and Growth*. Cambridge University Press.

Lynn, M. (1986) Determination and Quantification of Content Validity Index. *Nursing Research, 35*, 382-386. https://doi.org/10.1097/00006199-198611000-00017

Mann, W. (2007). German Deaf Children's Understanding of Referential Distinction in Written German and German Sign Language. *Education and Child Psychology, 24* (4), 57-75.

Marschark, M. (1993). *Psychological development of deaf children*. Oxford University Press.

Marschark, M., Shaver, D.M., Nagle, K.M., & Newman, L.A. (2015). Predicting the academic achievement of deaf and Hard-of-Hearing Students from individual, household, communication, and educational factors, *Exceptional Children, 81*(3), 350–369. https://doi.org/10.1177/0014402914563700

Matthews, S., & Yip, V. (2011). *Cantonese: A Comprehensive Grammar* (2nd ed.). Routledge.

Mayberry, R. I., Del Giudice, A. A., & Lieberman, A. M. (2011). Reading achievement in relation to phonological coding and awareness in deaf readers: A meta-analysis. *The Journal of Deaf Studies and Deaf Education*, *16*(2), 164-188.

Mayer, C., Trezek, B. J., & Hancock, G. R. (2021). Reading Achievement of Deaf Students: Challenging the Fourth Grade Ceiling. *The Journal of Deaf Studies and Deaf Education*, *26*(3), 427-437.

McBride-Chang, C. & Lin, Dan & Fong, Y.-C & Shu, Hua. (2012). Language and literacy development in Chinese children. *Oxford Handbook of Chinese Psychology*. **https://doi.org/10.1093/oxfordhb/9780199541850.013.0008**

McBride-Chang, C., Tong, X., Shu, H., Wong, A. M.-Y., Leung, K.-w., & Tardif, T. (2008). Syllable, phoneme, and tone: Psycholinguistic units in early Chinese and English word recognition. *Scientific Studies of Reading, 12*(2),171-194. **https://doi.org/10.1080/10888430801917290**

McBride-Chang, C., Wagner, R. K., Muse, A., Chow, B. W-Y., & Shu, H. (2005). The role of morphological awareness in children's vocabulary acquisition in English. Applied Psycholinguistics, 26(3), 415-435. **https://doi.org/10.1017/S014271640505023X.**

McDaniel, D., Cairns, H. S., & McKee, C. (Eds.). (1998). *Methods for Assessing Children's Syntax*. The MIT Press.

McGraw, K.O., & Wong, S.P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods, 1*, 30–46.

Meijer, R.R., & Nering, M.L. (1999). Computerized adaptive testing: overview and introduction. *Applied Psychological Measurement*, *23*(3), 187–194. **https://doi.org/10.1177/01466219922031310**

Min, S. & Aryadoust, V. (2021). A systematic review of item response theory in language assessment: Implications for the dimensionality of language ability. *Studies in Educational Evaluation, 68*, 1-10.

Moeller, M. P., Hoover, B., Putman, C., Arbataitis, K., Bohnenkamp, G., Peterson, B. & Stelmachowicz, P. (2007). Vocalizations of infants with hearing loss compared with infants with normal hearing: Part I–phonetic development. *Ear and Hearing*, *28*(5), 605-627.

Moog, J.S., & Geers, A.E. (1980). *Grammatical Analysis of Elicited Language – Complex Sentence Level.* Central Institute for the Deaf.

Moog, J.S., & Geers, A.E. (1985). *Grammatical Analysis of Elicited Language – Simple Sentence Level* (2nd ed.). St Louis, MO: Central Institute for the De, D. (2010). Partners in Progress: The 21st International Congress on Education of the Deaf and the Repudiation of the 1880 Congress of Milan. American Annals of the Deaf, 155(3), 309-310. Retrieved July 11, 2020, from **www.jstor.org/stable/26235069.**

Ng, H-Y. I. (2014). The Construction and Validation of the Cantonese Spoken Word Recognition Test (CanSWORT) to Measure Word Recognition Ability of Cantonese-speaking Population. [Unpublished PhD dissertation], The Chinese University of Hong Kong.

Nikolopoulos, T. P., Dyar, D., Archbold, S., & O'Donoghue, G. M. (2004). Development of Spoken Language Grammar Following Cochlear Implantation in Prelingually Deaf Children. *Archives of Otolaryngology-Head Neck Surgery*, *130*, 629-633.

Pan, D.J., Yang, X., Lui, K.F.H., LO, J.C.M., McBride, C., & Ho, C.S-H. (2021). Character and word reading in Chinese: Why and how they should be considered uniquely vis-à-vis literacy development, *Contemporary Educational Psychology, 65,* **https://doi.org/10.1016/j.cedpsych.2021.101961**.

Pinker, S. (1984). *Language learnability and language development.* Harvard University Press.

Qi, S., & Mitchell, R. E. (2012). Large-scale academic achievement testing of deaf and hard-of-hearing students: Past, present, and future. *Journal of deaf studies and deaf education*, *17*(1), 1-18.

Qian, D. (2002). Investigating the relationship between vocabulary knowledge and academic reading performance: An assessment perspective. *Language learning*, *52*(3), 513-536.

Quigley, S. P. (1969). Chapter VII: The Deaf and the Hard of Hearing. *Review of Educational Research*, *39*(1), 103–123. **https://doi.org/10.3102/00346543039001103**

Quigley, S. P., Wilbur, R. B., Power, D. J., Montanelli, D. S., & Steinkamp, M. W. (1976). *Syntactic Structures in the Language of Deaf Children*. U.S. Department of Health, Education, and Welfare, National Institute of Education.

Quigley, S. P. (1977). The language structure of deaf children. *Volta Review*, *79*(2), 73-84.

Quigley, S. P, Steinkamp, M., Power, D. & Jones, B. (1978). The assessment and development of language in hearing impairment individuals. *Journal of the Academy of Rehabilitative Audiology, 11*(1), 24-41.

Quigley, S. P., & Paul, P. V. (1994). *Language and Deafness* (2nd ed.). Singular Pub. Group.

Radford, A. (1990). The syntax of nominal arguments in early child English. *Language Acquisition*, *1*, 195-223.

Radford, A. (2004). *Minimalist Syntax: Exploring the structure of English*. Cambridge University Press.

Rasch, G. (1960). *Probabilistic Models for Some Intelligence and Attainment Tests.* Danish Institute for Educational Research.

Reise, S. P., Cook, K. F., & Moore, T. M. (2014). Evaluating the impact of multidimensionality on unidimensional item response theory model parameters. In S. P. Reise, & D. A. Revicki (Eds.), *Handbook of Item Response Theory Modeling: Applications to Typical Performance Assessment* (pp. 13–40). New York: Routledge.

Rinaldi, P., & Caselli, M. C. (2014). Language Development in a Bimodal Bilingual Child with Cochlear Implant: A Longitudinal Study. *Bilingualism: Language and Cognition, 17*, 798-809. **https://doi.org/10.1017/S1366728913000849**

Rogers, W. T., & Clarke, B. R. (1981). Psychometric characteristics of the Test of Syntactic Abilities Screening Test. *Educational and Psychological Measurement, 41*(1), 145-162.

Rinaldi, P., & Caselli, C. (2009). Lexical and grammatical abilities in deaf Italian preschoolers: The role of duration of formal language experience. *Journal of Deaf Studies and Deaf Education, 14*(1), 63-75.

Saidi, S.S., & Siew, N.M. (2019). Reliability and validity analysis of statistical reasoning test survey instrument using the Rasch measurement model. *International Electronic Journal of Mathematics Education, 14*(3), 535-546. **https://doi.org/10.29333/iejme/5755**.

Seymour, H., Roeper, T., & de Villiers, J. (2005). The DELV-NR (Norm-referenced version). The diagnostic evaluation of language variation. The Psychological Corporation.

Shrout, P.E., & Fleiss, J.L. (1979). Intraclass correlations: uses in assessing rater reliability. *Psychological Bulletin, 86,* 420-428.

Smith, F. (1994). *Understanding reading*. Hillsdale, N.J.: L. Erlbaum.

Spencer, P. E., & Marschark, M. (2010). *Evidence-based practice in educating deaf and hard-of-hearing students*. Oxford University Press.

Stinson, M. S., & Antia, S. (1999). Considerations in educating deaf and hard-of-hearing students in inclusive settings. *Journal of Deaf Studies and Deaf Education, 4*(3), 163-175.

Sze, F., Lo, C., Lo, L. & Chu, K. (2012). Lexical variations and diachronic change in Hong Kong Sign Language: preliminary observations. Paper presented in New Ways of Analyzing Variation in Asia Pacific 2, 1-4 August 2012, NINJAL, Tokyo.

Sze, F., Lo, C. Lo, L. & Chu, K. (2013). Historical development of Hong Kong Sign Language. *Sign Language Studies, 13*(2)*,* 155-185.

Sze, F., Wei, M. X., & Lam, D. (2020). Development of the Hong Kong Sign Language Sentence Repetition Test. *The Journal of Deaf Studies and Deaf Education*, *25*(3), 298-317.

Tager-Flusberg, H. (1994). Contributions to the field of language acquisition from research on atypical children. In H. Tager-Flusberg (Ed.), *Constraints on Language Acquisition: Studies of Atypical Children* (pp. 1-8). Lawrence Erlbaum Associates.

Takahashi, N., Isaka, Y., Yamamoto, T., & Nakamura, T. (2017). Vocabulary and grammar differences between deaf and hearing students. *The Journal of Deaf Studies and Deaf Education*, *22*(1), 88-104.

Tang, G., Li, Q., Li, J., Hu, Y-Y., Yiu, K-M., & Lam, D. (2020). *Chinese Grammatical Knowledge Assessment.* The Centre for Sign Linguistics and Deaf Studies, The Chinese University of Hong Kong.

Tang, G., Li, Q., Li, J., & Yiu, K-M., C. (2023) Chinese grammatical development of d/Deaf and hard-of-hearing children in a sign bilingualism and co-enrollment program. *American Annals of the Deaf, 167*(5), 675–699.

Tang, G., Yiu, C. K-M. & Lam, S. (2015). Awareness of HKSL and manually coded Chinese by deaf students in a Sign Bilingual and Co-enrollment Setting: A HK case study. In H., Knoors & M. Marschark (Eds). *Educating Deaf Learners: Creating a Global Evidence Base* (pp. 117-148). Oxford University Press.

Tang, G.W-L., Yiu, C.K-M., Lee, K.Y-S., Li, J., Li, Q., Ho, C. C-M. & Lam, D. C-F. (2022). Two languages are better than one: Establishing inclusive education for the Deaf and hard-of-hearing children in Hong Kong using a sign bilingualism and co-enrollment approach.

BrainChild, *21*(1), 9-30.

Taylor, I. (2002). Phonological Awareness in Chinese Reading. In: L.Wenling, J.S. Gaffney, & J.L. Packard (Eds.), *Chinese Children's Reading Acquisition.* Springer, Boston, MA. **https://doi.org/10.1007/978-1-4615-0859-5_3**.

The Hong Kong Society for the Deaf. (2009). *A Survey on the Difficulties and Challenges Encountered by Primary Students with Hearing Impairment in Integrated Education*.

The Linguistic Society of Hong Kong. (n.d.). *The Jyutping Scheme.* https://web.archive.org/web/20130426050642/http://www.lshk.org/node/47

Tong, X., McBride-Chang, C., Shu, H., & Wong, A. M. (2009). Morphological awareness, orthographic knowledge, and spelling errors: Keys to understanding early Chinese literacy acquisition. *Scientific Studies of Reading*, *13*(5), 426-452.

Traxler, C. B. (2000). The Stanford Achievement Test: National norming and performance standards for deaf and hard-of-hearing students. *Journal of Deaf Studies and Deaf Education*, *5*(4), 337-348.

T'Sou, B., Lee, T., Tung, P., Chan, A., Man, Y., & To, C. (2006). *Hong Kong Cantonese Oral Language Scale*. Hong Kong: City University of Hong Kong Press.

Tuller, L., & Jakubowicz, C. (2004). Développement de la morpho-syntacticxe du français chez des enfants sourds moyens. *Le Langage et l'Homme: Logopédie, 14*, 191-207.

United Nations (2016). Convention on the Rights of Persons with Disabilities and Optional Protocol. Retrieved from, **https://www.un.org/disabilities/documents/convention/convoptprot-e.pdf.**

Van Gent, T. (2016). Mental Health Problems of Deaf Children and Adolescents. In M. Marschark, V. Lampropoulou, & E.K. Skordilis (Eds.) *Diversity in Deaf Education*, (pp. 381-416). Oxford University Press.

Volpato, F. (2010). *The Acquisition of Relative Clauses and Phi-features : Evidence from Hearing and Hearing-impaired Populations*. [Unpublished doctoral dissertation], University of Venice.

Wang, Q., & Andrews, J. (2020). *Literacy and deaf education: toward a global understanding*. Gallaudet University Press.

Wang, Z. W., Lian, F. X., & Lin, Y. Q. (2018). On sentence patterns and syntactic errors in Intermediate-Grade hearing-impaired students' diaries. *Chinese Journal of Special Education. 216*(6), 35-40.

Watkin, P., McCann, D., Law, C., Mullee, M., Petrou, S., Stevenson, J., & Kennedy, C. (2007). Language ability in children with permanent hearing impairment: the influence of early management and family participation. *Pediatrics*, *120*(3), e694-e701.

Wiig, E.H., Semel, E., & Secord, W.A. (2013). *Determining the Severity of Language Disorder*. Pearson Education.

Wilbur, R., Goodhart, W., & Montandon, E. (1983). Comprehension of Nine Syntactic Structures by Hearing-Impaired Students. *The Volta Review*, 328-345.

Wilson, M. (2005). *Constructing Measures*: *An Item Response Modeling Approach*. Erlbaum.

Wong, A. M.-Y., Leung, C., Ng, A. K.-H., Cheung, P. S.-P., Siu, E. K.-L., To, K.-S., Sam, S. K.-L., Cheung, H.-T., Lo, S.-K., & Lam, C. C.-C. (2019). *Technical Manual of the Hong*

Kong Test of Preschool Oral Language (Cantonese) (TOPOL). Department of Health, Hong Kong SAR Government.

Wong, R.S.M., Ho, F.K.W., Wong, W.H.S., Tung, K.T.S., Chow, C.B., Rao, N., Chan, K.L., & Ip, P. (2018). Parental involvement in primary school education: its relationship with children's academic performance and psychosocial competence through engaging children with school. *Journal of Child and Family Studies, 27,* 1544–1555. **https://doi.org/10.1007/s10826-017-1011-2**

World Federation of the Deaf (2018). *WFD Position Paper on Inclusive Education.* Retrieved from: **https://wfdeaf.org/news/resources/5-june-2018-wfd-position-paper-inclusive-education**

Wright, B. D., & Linacre, J. M. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions, 8,* 370-371.

Wright, B.D., & Stone, M.H. (1979). *Best Test Design.* MESA.

Wright, B.D., & Stone, M.H. (1999). *Measurement Essentials.* (2nd ed.). Range Inc.

Wu, X., Anderson, R. C., Li, W., Wu, X., Li, H., Zhang, J., Zheng, Q., Zhu, J., Shu, H., Jiang, W., Chen, X., Wang, Q., Yin, L., He, Y., Packard, J., & Gaffney, J. S. (2009). Morphological Awareness and Chinese Children's Literacy Development: An Intervention Study. *Scientific Studies of Reading, 13*(1), 26–52.

https://doi.org/10.1080/ 10888430802631734.

Yamashita, J. (1999). *Reading in a First and a Foreign Language: A Study of Reading Comprehension in Japanese (the L1) and English (the L2)* [Doctoral dissertation], University of Lancaster.

Yan, X., Cheng, L., & Ginther, A. (2019). Factor analysis for fairness: Examining the impact of task type and examinee L1 background on scores of an ITA speaking test. *Language Testing, 36*(2), 207–234. **https://doi-org.easyaccess1.lib.cuhk.edu.hk/10.1177/ 0265532218775764.**

Yip, P-C., & Rimmington, D. (2004). *Chinese: A Comprehensive Grammar.* New York: Routledge.

Yiu, K.-M. (2004) Acquisition of restrictive relative clauses by orally–trained profoundly hearing impaired children. [Unpublished master's thesis], The Chinese University of Hong Kong.

Yiu, K. M. (2012). Acquisition of Cantonese passive *bei2* constructions by deaf children. [Unpublished master's thesis], The Chinese University of Hong Kong.

Yiu, C. K. M., Tang, G., & Ho, C. C. M. (2019). Essential ingredients for sign bilingualism and co-enrollment education in the Hong Kong context. In M. Marschark, S. Antia, & H. Knoors (Eds.), *Co-Enrollment in Deaf Education* (pp. 83-106), New York: Oxford University Press.

Zwick, R., Thayer, D.T., & Lewis, C. (1999). An empirical Bayes approach to Mantel-Haenszel DIF analysis. *Journal of Education measurement, 36*(1), 1-28.

Appendices

Appendix A : The 172 items of the Chinese Grammatical Assessment and their Respective Grammatical Categories and Sub-categories

Morphosyntactic Categories	Sub-Categories[a]	No. of Items	Task Type[b]
S01 *Ba*-construction	1.1 With bare verb	4	GJ
	1.2 With complement phase	4	GJ
	1.3 Basic *Ba*-construction	4	TVJ
S02 Passive	2.1 Long passives	4	TVJ
	2.2 Short passives	4	TVJ
S03 Binding	3.1 Complex reflexive	4	PS
	3.2 Simple reflexive	4	PS
	3.3 Pronoun	4	PS
	3.4 Pronoun with verb 幫 'help'	4	PS
	3.5 Reflexive with verb 幫 'help'	4	PS
S04 Relative clause	4.1 Relative clause - Subject-Object (SO)	4	PS
	4.2 Relative clause - Subject-Subject (SS)	4	PS
	4.3 Relative clause - Object-Object (OO)	4	PS
	4.4 Relative clause - Object-Subject (OS)	4	PS
S05 Comparative	4.1 不比 'not-compare' - less	2	TVJ
	4.2 不比 'not-compare' - more	2	TVJ
	4.3 不比 'not-compare' - same	2	TVJ
	4.4 沒有 'no'	4	TVJ
	4.5 Basic comparative 比 'compare'	4	PS
S06 Quantification	6.1 All	4	TVJ
	6.2 Every	4	TVJ
	6.3 Negator-quantifier	4	PS
	6.4 Quantifier-negator	4	PS
S07 Double-object construction	7.1 Basic double-object construction	4	MC
S08 Locative existential	8.1 Animate subject	2	MC
	8.2 Inanimate subject	2	MC
S09 Control	9.1 Object control	4	PS
S10 Cleft sentences	10.1 Which place/What time	4	MC
S11 Question	11.1 Question word - modal	4	MC
	11.2 Modal - question word	4	MC
S12 Morpheme distinction	12.1 Particle 的 (*dik1*)	4	MC
	12.2 Particle 地 (*dei6*)	4	MC
	12.3 Particle 得 (*dak1*)	4	MC

Morphosyntactic Categories	**Sub-categories**[a]	**No. of Items**	**Task Type**[b]
S13 Negation	13.1 Negator 不 'not'	4	MC
	13.2 Negator 沒有 'no'	4	MC
S14 Preposition	14.1 Preposition – 對 (*deoi3*)	2	MC
	14.2 Preposition – 跟 (*gan1*)	2	MC
	14.3 Preposition – 從 (*cung4*)	2	MC
	14.4 Preposition – 向 (*hoeng3*)	2	MC
	14.5 Preposition – 在 (*zoi6*)	2	MC
S15. Localizer	15.1 With localizer	4	GJ
	15.2 Without localizer	4	GJ
S16 Aspect	16.1 Perfective	4	TVJ
	16.2 Progressive	4	TVJ
S17 Question word	17.1 *wh*-adjunct	4	MC
	17.2 *wh*-argument	4	MC
S18 Question particle	18.1 A-not-A with particle 嗎 (*maa1*)	4	GJ
	18.2 A-not-A with particle 呢 (*ne1*)	4	GJ
	Total no. of items:	**172**	

[a] For the terminology in written Chinese, the gloss in English is provided such as 幫 'help' and in some cases, especially for some function words with multiple meanings or no direct meaning, a phonetic representation following Cantonese Jyutping romanization system (The Linguistic Society of Hong Kong, n.d) are provided for readers' reference, for example, 呢 (*ne1*).

[b] GJ=Grammaticality Judgement; MC=Multiple Choice; PS=Picture Selection; TVJ=Truth Value Judgement

Appendix B: Platform for Expert Panel’s Content Validation of CGA (the Chinese Version and the Translated English Version)

中文語法評估 (CGA)

專業評審網頁

再次感謝你撥冗參與「中文語法評估」(Chinese Grammatical Assessment- CGA) 的專業評審工作。整個評審工作主要分爲三個部份: 第一部份是關於檢視評估工具的整體運作模式；第二部份是各題目的設計，這些題目與評估目標的相關性和代表性；而最後一部份是收集你對這個評估工具的整題意見和建議。

各專家顧問提供的意見會在網上平台自動記錄下來，由於要檢視的題目比較多，你可以分開幾次填寫當中的問卷，直至完成所有題目，提交給我們爲止。在評審期間你亦可以隨時更改你的答案。你所提供的寶貴意見可以幫助我們重新檢視這評估工具的題目，讓我們可以挑選最好的題目成爲最後的版本，更有效地測量聾、健學童的中文語法能力。

爲方便記錄各位專家顧問的意見和進行跟進，在進行各評審工作前，請先簡單填寫你的個人資料。在完成所有評審後，我們會總結所有專家顧問的意見作出綜合報告，若在報告中需要引用個別專家顧問的意見，都會以匿名方式處理的。另外，你可以隨時更改你的個人資料和評審意見，亦可以隨時退出參與這評審的工作。

個人資料

英文姓名: ______________________

專業範疇: 言語治療師/小學中文老師/其他專業(請填寫): ____.

以上專業工作的經驗: _______________年

你有服務聾童或弱聽學童的經驗嗎？ 有/沒有

若有，請問約有多少年？_______________年

第一部份:：評估工具的整體運作模式

在這部分，我們會逐一介紹「中文語法評估」的運作模式，然後邀請專家評審就着每一個環節的設計進行檢視及提供意見。

評估簡介

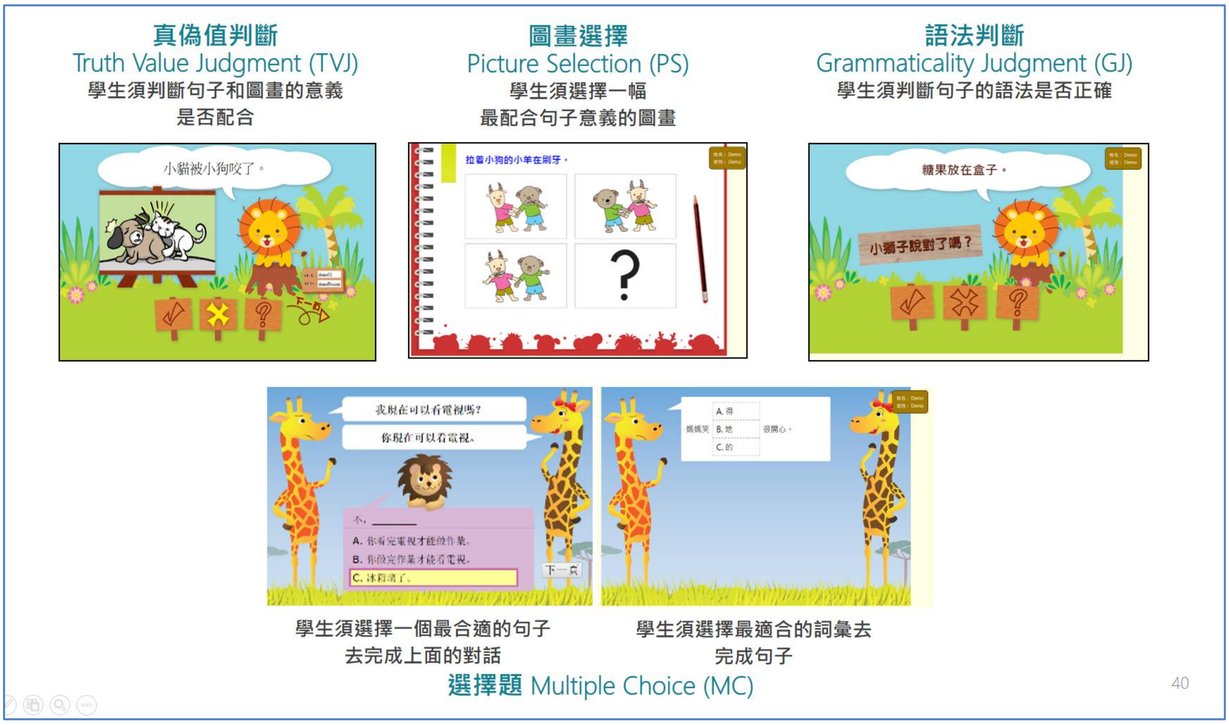
「中文語法評估(CGA)」 總共有 172 題題目，當中透過 48 種不同句式來評估學童對 18 個中文語法知識範疇的理解能力 。根據小學生程度，題目長度限於 5 到 12 個字組成 (有關 18 個語法知識項目請看圖表一)。

編號	語法項目		例句
S01	Ba-construction	把字句	小明把花瓶打破了。
S02	Passives	被動句	花瓶被小明打破了。
S03	Binding	約束句	小明的哥哥在畫他。
S04	Relative clauses	關係從句	戴著帽子的男孩在踢球。
S05	Comparatives	比較句	小明比小華高。
S06	Quantification	量化句	所有男孩都在畫畫。
S07	Double-object Construction	雙賓句	小明送給老師一束花。
S08	Locative Existential	處所句	操場上站著一個男孩。
S09	Control	兼語句	小明要姐姐講故事。
S10	Cleft Sentences	分裂句	小明是後天參加圍棋比賽的。
S11	Question	疑問詞及情態動詞	媽媽怎麼會去學校？
S12	Morpheme Distinction	結構助詞	小明笑得很開心。
S13	Negation	否定句	小明昨天沒有參加圍棋比賽。
S14	Preposition	介詞	小明向公園跑去。
S15	Localizer	方位詞	小明坐在沙發上。
S16	Aspect	體貌詞	小明喝了一杯水。
S17	Question words	疑問詞	小明什麼時候參加圍棋比賽？
S18	Question Particles	疑問語氣詞	小明要不要參加圍棋比賽呢？

圖表一：「中文語法評估(CGA)」 中 18 個語法知識項目

A) 介紹不同題目回答方式的錄像

這評估工具以四類型題目進行測試，包括: 圖畫選擇(Picture Selection)、真偽值判斷(Truth Value Judgement)、語法判斷(Grammaticality Judgement) 和選擇題 (Multiple Choice)（請參閱圖表二）。



真偽值判斷
Truth Value Judgment (TVJ)
學生須判斷句子和圖畫的意義是否配合

小貓被小狗咬了。

圖畫選擇
Picture Selection (PS)
學生須選擇一幅最配合句子意義的圖畫

拉著小狗的小羊在刷牙。

語法判斷
Grammaticality Judgment (GJ)
學生須判斷句子的語法是否正確

糖果放在盒子。

小獅子說對了嗎？

我現在可以看電視嗎?

你現在可以看電視。

不，______

A. 你看完電視才能做作業。

B. 你做完作業才能看電視。

C. 冰箱壞了。

媽媽笑 A. 得 B. 地 C. 的 很開心。

學生須選擇一個最合適的句子去完成上面的對話

學生須選擇最適合的詞彙去完成句子

選擇題 Multiple Choice (MC)

圖表二：「中文語法評估(CGA)」中的四類題目

「中文語法評估」以網上評估模式進行。由於這評估的對象包括所有聾/弱聽和健聽的學生，爲統一給予指示的模式和內容，盡量減低聽力障礙的影響，在評估前所有參與學童都須先觀看一段答題指示的錄像，讓學童了解評估中不同類型題目的回答方式。以下是該錄像，請檢視這錄像的內容及指示能否清晰讓學生明白四類評估題目的回答方法。

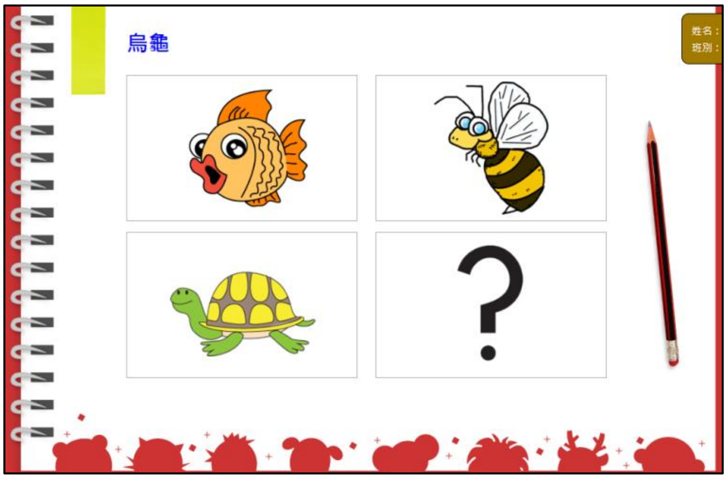
Video Instructions

B) 例題和答題方式

在評估開始前，評估系統會先給予學童 2 題 例題，讓學生嘗試回答，熟習答題方式。另外，在整個評估過程中，所有題目不須任何書寫，學童只須在電腦版面上點按最好的答案就可以了。另外，爲減低學習效應 (learning effects)，這評估會以隨機抽選的題目次序進行，故此，每一位學生都會以不同題目次序接受評估的。

C) 詞彙測試

此評估的對象是香港小一到小六學生，除典型發展的學童外，也包括聾童和弱聽學童。爲確定學生已經掌握當中的詞彙，在設計題目時會盡量運用學前或初小常用的詞彙。這些詞彙會重複在題目中使用。在正式進行語法評估之前，學生都需要先進行詞彙評估，以了解學生是否都掌握這些詞彙，以致我們可以確定評估結果不受學生詞彙能力影響，評估結果能真正反映學生的中文語法知識。這詞彙評估有 32 題，以四選一形式進行(請參閱圖表三及四)。



圖表三: 詞彙評估題目

名詞			動詞		形容詞
叉	白兔	尺子	掃地	踢	傷心
圖書館	烏龜	白紙	爬樹	穿	長
屋頂	獅子	課本	騎馬	推	瘦
草地	蜜蜂	箱子	畫畫	抱	
衣架	飛機	籃球	跳繩	敲	
衣服	火車		睡覺		

圖表四：32 個在「中文語法評估」中常用詞彙

根據以上的簡介，請專家顧問根據以下題目給予你的評價，評分方式如下:

1 = 非常不合適　　2 = 不太合適　　3 =一般　　4 = 頗合適　　5 =非常合適

	「中文語法評估」的評估模式	各項設計的合適程度	對這方面的建議(如有)
1	以網上平台方式進行整個評估	□1 □2 □3 □4 □5	
2	以錄像方式向聾、健兒童介紹不同題目的回答方法	□1 □2 □3 □4 □5	
3	錄像的內容和表達方法	□1 □2 □3 □4 □5	
4	進行語法評估前先給予詞彙評估的安排	□1 □2 □3 □4 □5	
5	詞彙評估的數量	□1 □2 □3 □4 □5	
6	正式進行前先嘗試做例題的安排	□1 □2 □3 □4 □5	
7	題目以電腦抽選方式進行，令每次評估時題目都以不同次序出現	□1 □2 □3 □4 □5	
8	學童回答後仍然可以更改自己的選擇	□1 □2 □3 □4 □5	

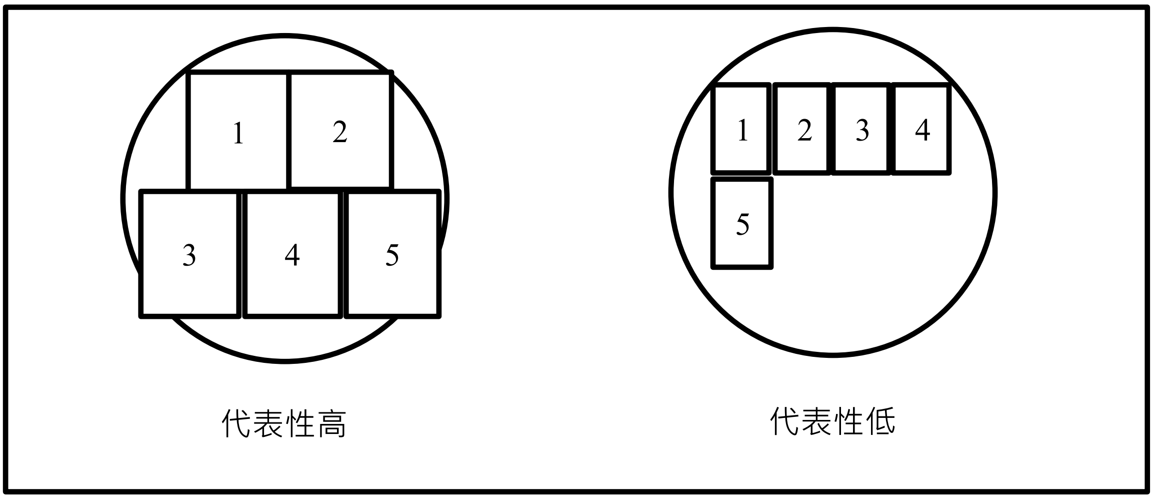
如對「中文語法評估」評估的運作模式有任何其他意見，請在下面填寫：

第二部份

這部份是檢視「中文語法評估」中選取的中文語法知識項目，與及所包含的內容是否有足夠的代表性(representativeness)，另外，因應這些擬定的內容而設計的題目又能否真正反映學生的中文語法知識，換句話說，是檢視題目設計與評估目標是否有清晰的相關性(relevance)。以下請專業顧問先檢視評估內容的代表性。

評估內容的代表性(representativeness)

評估內容的「代表性」是發展評估工具時非常重要的一個環節，要關注的是評估工具中選取的內容重點能否涵概最重要和最有代表性的評估範疇（facets），以反映評估對象的潛在特質(latent trait)。故此，專家顧問可以就着現時「中文語法評估」中選取的 18 個主要語法知識項目去檢視一下這些評估內容的代表性(請參與圖表四)，在當中是否已涵概對小學生而言最主要的語法知識項目，無須再加其他題目。如代表性低，表示評估的語法知識範疇不足，需要增加更多不同語法知識範疇的題目。



圖表四: 評估題目的代表性

以下請專家顧問根據以下圖表中列出的 「中文語法評估」不同語法知識項目、測試重點和例子，檢視一下在這評估裏所選取的語法知識項目是否有足夠的代表性。評分方式如下:

1 = 代表性非常低　2 = 代表性低 3 = 代表性一般　4 = 代表性高　5 = 代表性非常高

編號	語法項目		例句	測試重點	這項語法知識的代表性	其他意見(如有)
S01	Ba-construction	把字句	小明把花瓶打破了。	把字句的不同句型及意思	□1 □2 □3 □4 □5	
S02	Passives	被動句	花瓶被小明打破了。 小明被打了。	長、短被動句的句型及意思	□1 □2 □3 □4 □5	
S03	Binding	約束句	小明的哥哥在畫他。	約束語“自己”跟代名詞“他”的分別	□1 □2 □3 □4 □5	
S04	Relative clauses	關係從句	戴著帽子的男孩在踢球。	關係子句的理解，尤其是句子中各參與者的關係	□1 □2 □3 □4 □5	
S05	Comparatives	比較句	小明比小華高。 西瓜不比蘋果甜。	基本比較句，與及當比較詞和否定詞同時出現時的意思	□1 □2 □3 □4 □5	
S06	Quantification	量化句	所有男孩都在畫畫。	全稱量詞“所有”、“有些”、“每”的分別與用法	□1 □2 □3 □4 □5	
S07	Double-object Construction	雙賓句	小明送給老師一束花。	雙賓句的語序	□1 □2 □3 □4 □5	
S08	Locative Existential	處所存在句	操場上站著一個男孩。	以地點作為主語的句子，並明白此類句型與典型陳述句的分別	□1 □2 □3 □4 □5	
S09	Control	控制句 兼語句？	小明要姐姐講故事。	控制句的意思及語法特性	□1 □2 □3 □4 □5	
S10	Cleft Sentences	分裂句	小明是後天參加圍棋比賽。	帶“是……的”分裂句的意思及語法特性	□1 □2 □3 □4 □5	
S11	Question & Modal	疑問句	媽媽怎麼會去學校？	疑問詞和情態動詞同時出現時的意思	□1 □2 □3 □4 □5	
S12	Morpheme Distinction	結構助詞	小明笑得很開心。	“的”、“地”、“得”這三個結構助詞的分別與用法	□1 □2 □3 □4 □5	
S13	Negation	否定詞	小明昨天沒有參加圍棋比賽。	否定詞“不”和“沒有”的分別與用法	□1 □2 □3 □4 □5	
S14	Preposition	介詞	小明向公園跑去。	介詞“對”、“跟”、“從”、“向”、“在”的分別與用法	□1 □2 □3 □4 □5	
S15	Localizer	方位詞	小明坐在沙發上。	方位詞（如：“上、裡”）的語法特性	□1 □2 □3 □4 □5	
S16	Aspect	體貌詞	小明喝了一杯水。	體貌詞“在”和“了”的分別	□1 □2 □3 □4 □5	
S17	Question words	疑問詞	小明什麼時候參加圍棋比賽？	不同疑問詞 (如:“誰”、“什麼”或“好不好”)的分別與用法	□1 □2 □3 □4 □5	
S18	Question Particles	疑問語氣詞	小明要不要參加圍棋比賽呢？	不同疑問語氣詞（如：“嗎、呢”）的語法特性	□1 □2 □3 □4 □5	

1. 根據小學生的中文讀寫發展，你認爲所選取的 18 項中文語法知識項目是否有足夠的代表性呢？

□ 代表性非常低　　代表性低　　□ 代表性一般　　□ 代表性高　　□代表性非常高

2. 你覺得有哪些語法知識項目需要刪減？

____

____

______

3. 有需要增加其他語法知識項目的需要嗎？請建議。

____

____

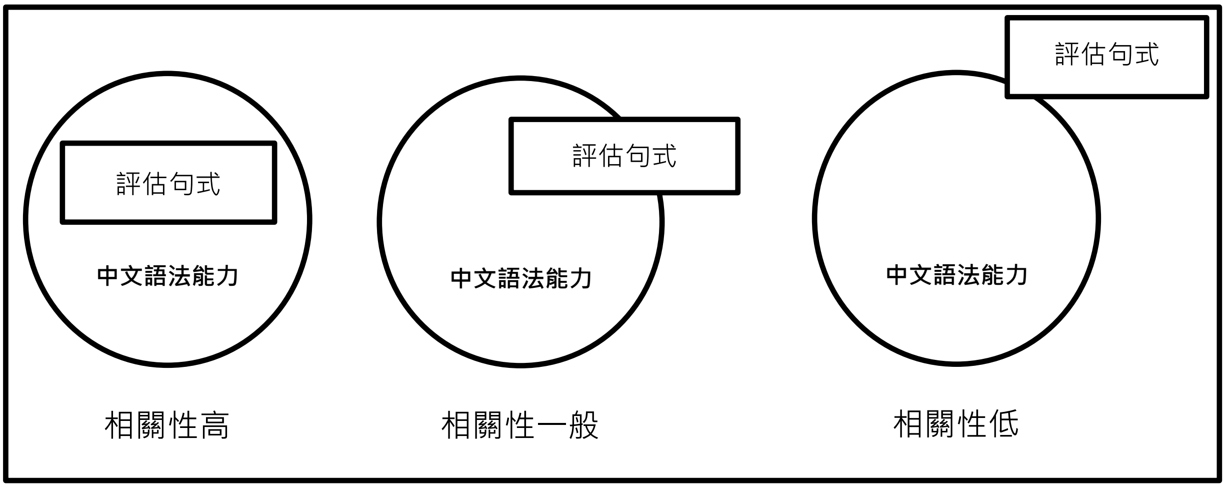
_____

4. 如有其他任何意見，請在下面提供。

題目的相關性(relevance)

「相關性」 是指這些用作評估的題目能否適切地反映評估對象的目標能力，以「中文語法評估」來說，就是學童的中文語法能力。相關性跟三方面有關，(一) 評估的目標；(二) 評估背後的理論; 和 (三) 評估題目包含的元素(包括題目內容及表達形式、答案選項、答題方式等)都能針對中文語法來作測試的。中文語法能力是由不同層面(facet)的語法知識所組成的，例如在「妹妹吃了飯」這個句子中「了」是體貌詞，它在句子中可以修飾動詞，帶來「完成」的時態，是中文語法中一個重要的語法知識項目 (Huang, Li, & Li, 2009)，也是兒童語言獲得研究常見的重點。故此，相關性高的題目就是那些可以清晰評核學生中文語法知識的題目。體貌詞「了」是中文語法當中其中一個重要的評估項目，與本身評估目標有很高相關性。若使用同一

句子「妹妹吃了飯」作爲評估題目，但評估重點只放在「吃飯」這個動詞的概念，那評估結果與中文語法能力這個評估目標的 關係就比較薄弱，相關性也就比較低了(請參與圖表五)。



圖表五: 評估題目的相關性

在下面我們會把「中文語法評估」中 172 題題目根據不同類型題目列出來，請專家顧問根據以下評分方式檢視這些題目與小學生中文語法能力的相關性*：

1 = 相關性非常低　2 = 相關性低　3 = 相關性一般　4 = 相關性高

5 = 相關性非常高

除此以外，也請檢視一下每題題目設計的合適性，即題目能否有效評估目標語法知識#:

1 = 非常不合適　2 = 不合適　3 =一般　4 = 合適

5 = 非常合適

（按 4 類題目，然後按語法項目分開逐題題目列出來）

語法知識項目	題目	所有答案選項（標示正確答案）	與中文語法能力的相關性*	題目的設計是否合適#	對測試題目的其他意見（如有）
比較句		哥哥比姐姐瘦。	□1 □2 □3 □4 □5	□1 □2 □3 □4 □5	
		弟弟要求妹妹話一幅畫。	□1 □2 □3 □4 □5	□1 □2 □3 □4 □5	
		蟲子從洞裏爬出來。	□1 □2 □3 □4 □5	□1 □2 □3 □4 □5	

		海龜在海里游。	□1 □2 □3 □4 □5	□1 □2 □3 □4 □5	
		小羊抱著的小狗在睡覺。	□1 □2 □3 □4 □5	□1 □2 □3 □4 □5	
		騎著小羊的小狗在唱歌。	□1 □2 □3 □4 □5	□1 □2 □3 □4 □5	

第三部分

整體而言，你對於「中文語法評估」有甚麼評價呢？以下希望你可以給予我們一些整體上的意見，以幫助我們改善整個評估的設計和用作評估的題目。

1) 你認爲「中文語法評估」的名稱合適嗎？

□ 非常不合適 □不合適 □ 一般 □ 合適 □非常合適

2) 你認爲「中文語法評估」的評核方法合適嗎？

□ 非常不合適 □不合適 □ 一般 □ 合適 □非常合適

3)「中文語法評估」的題目適合用作評估典型發展小學生的中文語法能力嗎？

□ 非常不合適 □不合適 □ 一般 □ 合適 □非常合適

4) 「中文語法評估」的題目適合用作評估聾生和弱聽小學生的中文語法能力嗎？

□ 非常不合適 □不合適 □ 一般 □ 合適 □非常合適

5) 整體而言，你對「中文語法評估」有其他建議嗎？請在下面填寫。

Reference:

Huang, C.-T. J., Li, Y.-H. A., & Li, Y. (2009). *The Syntax of Chinese*. Cambridge University Press.

Chinese Grammar Assessment (CGA)
Webpage for Expert Review

Thank you again for taking the time to participate in the professional review of the Chinese Grammatical Assessment (CGA). The whole assessment process is mainly divided into three parts: the first part is to review the overall operation of the assessment tool; the second part is to review the relevance and representativeness of the items according to the assessment objectives; and the final part is to gather your comments and suggestions on the assessment tool.

The opinions provided by each consultant will be automatically recorded on the online platform. Since there are many questions to review, you can fill in the questionnaire in several times until you have completed all the questions and submitted them to us. You can also change your answers at any time during the assessment period. Your valuable input can help us review the items of this assessment tool, so that we can select the best items for the final version and asses the Chinese grammatical knowledge of the typically developing and deaf or hard-of-hearing (DHH) students more effectively.

In order to record down all the opinions from our consultants and follow-up on the matters, please simply fill in your personal information before proceeding with the review. After all the parts have been reviewed, we will summarize the opinions from all consultants for a consolidated report, and if the opinions of individual consultants are quoted in the report, they will be treated anonymously. In addition, you can change your personal information and assessment comments at any time. You can also withdraw from participating in this review at any time.

Personal Data

English name: ____________

Professional category: Speech therapist / Primary Chinese teacher / Other professions (please indicate): ____

Years of experience in the above professional work:__________________________________

Do you have any experience working with deaf or hard-of-hearing students? Yes/no

If so, how many years? __________

Part I: <u>Review on the overall operation of the assessment tool</u>

In this section, we will introduce the operation elements of the Chinese Grammatical Assessment, and then invite you, as our consultants, to review the operations and provide opinions on their design.

<u>Introduction to assessment</u>

The Chinese Grammar Assessment (CGA) has a total of 172 test items, of which 48 different grammatical categories are used to assess the students' understanding of the Chinese grammatical knowledge. Depending on the level of primary school students, the length of the items is limited to 5 to 12 words (see the 18 Chinese Grammatical Categories in Table 1 below).

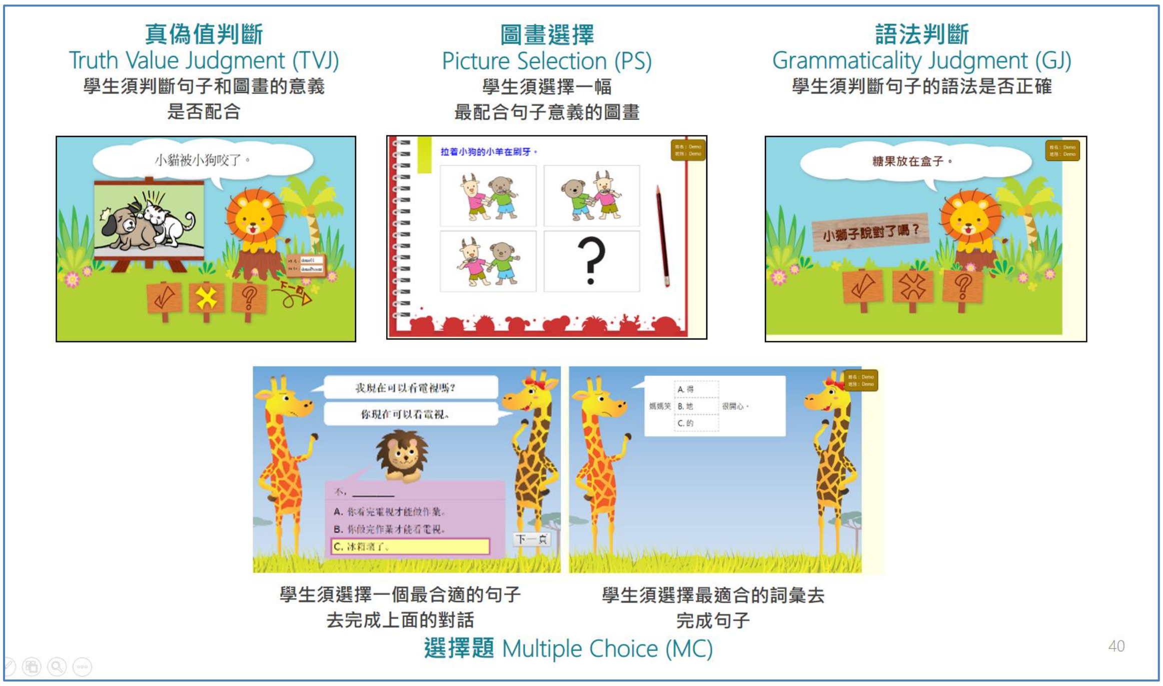
Table 1: The 18 grammatical categories included in the Chinese Grammar Assessment (CGA).

Category	**Grammatical Category**		**Examples**
S01	Ba-construction	把字句	小明把花瓶打破了。 'Siu Ming broke the vase.'
S02	Passives	被動句	花瓶被小明打破了。 'The vase was broken by Siu Ming.'
S03	Binding	約束句	小明的哥哥在畫他。 'Siu Ming's brother is painting him.'
S04	Relative clauses	關係從句	戴著帽子的男孩在踢球。 'The boy in a hat is playing football.'
S05	Comparatives	比較句	小明比小華高。 'Siu Ming is taller than Siu Fa.'
S06	Quantification	量化句	所有男孩都在畫畫。 'All the boys were drawing.'
S07	Double-object Construction	雙賓句	小明送給老師一束花。 'Siu Ming gave the teacher a bouquet of flowers.'
S08	Locative Existential	處所句	操場上站著一個男孩。 'There is a boy standing in the playground.'
S09	Control	兼語句	小明要姐姐講故事。 'Siu Ming asked his sister to tell a story.'
S10	Cleft Sentences	分裂句	小明是後天參加比賽的。

			‘Siu Ming will participate in a competition the day after tomorrow.’
S11	Question & Modal	疑問詞及情態動詞	媽媽怎麼會去學校？ ‘How did mom go to school?’
S12	Morpheme Distinction	結構助詞	小明笑得很開心。 ‘Siu Ming smiled happily.’
S13	Negation	否定句	小明昨天沒有參加比賽。 ‘Siu Ming did not participate in the competition yesterday.’
S14	Preposition	介詞	小明向公園跑去。 ‘Siu Ming is running towards the park.’
S15	Localizer	方位詞	小明坐在沙發上。 ‘Siu Ming is sitting on the sofa.’
S16	Aspect	體貌詞	小明喝了一杯水。 ‘Siu Ming has drunk a glass of water.’
S17	Question words	疑問詞	小明什麼時候參加比賽？ ‘When does Siu Ming participate in the competition?’
S18	Question Particles	疑問語氣詞	小明要不要參加比賽呢？ ‘Does Siu Ming want to participate in the competition?’

A) Videos introducing the answers to different questions

There are four different tasks used in this assessment tool, including Picture Selection and Truth Value Judgement, Grammaticality Judgement and Multiple Choice (see Figure 1).



真偽值判斷
Truth Value Judgment (TVJ)
學生須判斷句子和圖畫的意義是否配合

圖畫選擇
Picture Selection (PS)
學生須選擇一幅最配合句子意義的圖畫

語法判斷
Grammaticality Judgment (GJ)
學生須判斷句子的語法是否正確

學生須選擇一個最合適的句子去完成上面的對話

學生須選擇最適合的詞彙去完成句子

選擇題 Multiple Choice (MC)

40

Figure 1: Four different tasks used in the Chinese Grammar Assessment (CGA).

The Chinese Grammar Assessment is administered as an online assessment. The targets of this assessment includes DHH and typically developing (TD) students. To unify the instructions for the students, a video was used to brief the students how they should work on the different tasks so as to reduce their barriers to comprehending the instructions. Below is the video. Please help to review if the contents and instructions are clearly explained in the video.

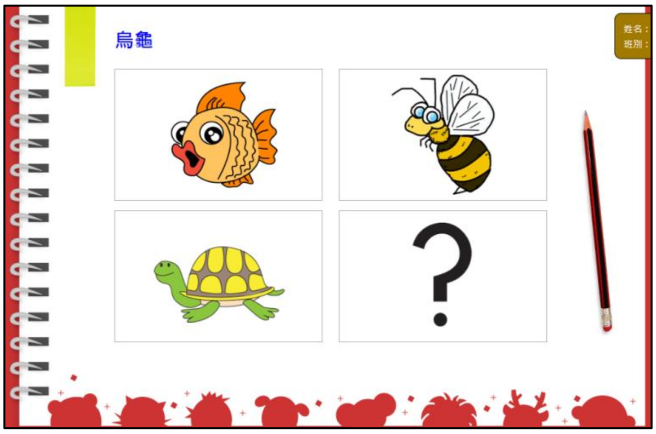
Video Instructions

B) Sample questions and answering methods

Before the assessment begins, the assessment system will give the student two sample questions for them to familiarize themselves with the assessment methods. In addition, during the whole assessment process, all students only need to click the best answer on the computer. They do not need to write anything. In addition, in order to reduce the learning effects, the questions will be randomly selected for the students. Therefore, each student will be assessed in a different order of questions.

C) Vocabulary test

This assessment targets Primary One to Primary Six students in Hong Kong including students with typical development and also Deaf and Hard-of-hearing students. To ensure that students have already mastered the principle vocabularies, vocabularies that are frequently used in preschool or primary school setting are adopted when designing the questions, and they will be used repeatedly in the test items. A vocabulary test will be conducted before the grammatical assessment to see if the students have already mastered the major vocabularies used in CGA. This is to ensure that the assessment results can truly reflect the students' Chinese grammatical knowledge. The results will not be affected by their vocabulary knowledge. The test is conducted in multiple choice questions with four answer choices (see Figure 2 and Table 2).



烏龜

Figure 2: Test item for vocabulary assessment

Table 2: Thirty-two frequently used vocabularies in CGA

noun			verb		adjective
fork	rabbit	ruler	sweep the floor	kick	sad
library	turtle	White paper	climb the trees	wear	long
roof	lion	textbook	riding a horse	push	thin
grassland	bee	box	draw	embrace	
hangers	plane	basketball	rope jumping	knock	
clothes	train		Sleep		

According to the above introduction, please give your evaluation based on the following questions and mark your scores according to the following scheme:

1 = very poor relevance 2 = poor relevance 3 = fair relevance
4 = good relevance 5 = very good relevance

	Operational elements of CGA	The relevance of each item	Recommendations in this regard, if any
1	Operating as a web-based online assessment.	□1 □2 □3 □4 □5	
2	Displaying items randomly by the computer - every time in a different order.	□1 □2 □3 □4 □5	
3	Students can change their answers before their submission.	□1 □2 □3 □4 □5	
4	Using an animated video to explain how to answer the different types of questions.	□1 □2 □3 □4 □5	
5	The contents and the illustration of the video.	□1 □2 □3 □4 □5	
6	Doing trial items before doing the test items.	□1 □2 □3 □4 □5	
7	Receiving a vocabulary test before doing CGA.	□1 □2 □3 □4 □5	
8	The number of words in the vocabulary test.	□1 □2 □3 □4 □5	

If you have any other comments on how the Chinese grammatical assessment should be operated, please fill in space below:

Part II:

This part examines whether the content covered by the Chinese grammatical categories selected for CGA are of good representativeness. In addition, whether the test items designed according to these proposed contents can truly reflect the student's knowledge of Chinese grammar. In other words, the aim is to check if the design of the items are relevant to the assessment objectives. You are invited to check the following assessment contents.

Representativeness of the Assessment Contents

The representativeness of assessment content is essential in the development of an assessment tool. Attention should be given to ensure that the contents are targeted on the latent trait of the testees, covering the major contents and the most representative facets. Therefore, expert panels can check with the representativeness of the 18 Chinese grammatical categories, to see if they are having good representativeness for the grammatical knowledge required for primary school students (please refers to Table 4). If the current grammatical categories have already covered the major grammatical knowledge, no other categories are required. Low representativeness indicates that grammatical domain assessed is insufficient. More grammatical categories should be included.

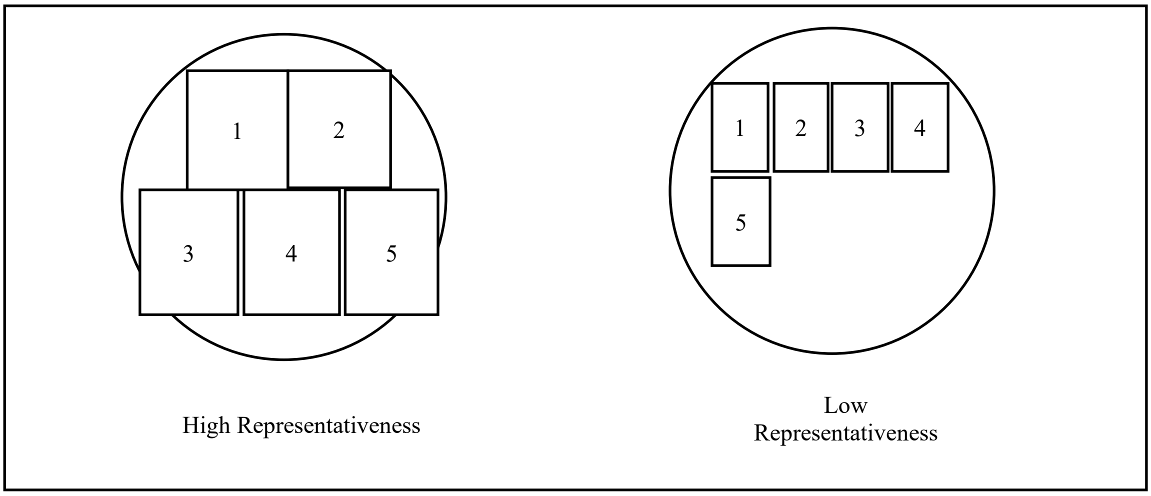


Figure 5: The representativeness of the grammatical categories

In the table below, all the different grammatical categories, their focus of assessment and the examples were listed out. Please review on the representativeness of these grammatical categories and provide your scores. The scoring system is as follows:

1 = very poor representativeness 2 = poor representativeness 3 = fair representativeness
4 = high representativeness 5 = very high representativeness

Code	Grammatical categories		Sample sentences	Assessment focus	Representativeness of the category	Other comments (if any)
S01	*ba*-construction	把字句	小明把花瓶打破了。 'Siu Ming broke the vase.'	The different types of *ba*-constructions and their respective semantics	□1 □2 □3 □4 □5	
S02	Passives	被動句	花瓶被小明打破了。 'The vase was broken by Siu Ming.'	The structures and meanings of long and short passive constructions	□1 □2 □3 □4 □5	
S03	Binding	約束句	小明的哥哥在畫他。 'Siu Ming's brother is painting him.'	The difference between the semantics of reflexive pronouns and personal pronouns	□1 □2 □3 □4 □5	
S04	Relative clauses	關係從句	戴著帽子的男孩在踢球。 'The boy in a hat is playing football.'	The understanding of different types of relational clauses, especially the relationships between the subjects and objects	□1 □2 □3 □4 □5	
S05	Comparatives	比較句	小明比小華高。 'Siu Ming is taller than Siu Fa.'	The semantics of basic comparatives and the comparatives with different negators	□1 □2 □3 □4 □5	
S06	Quantification	量化句	所有男孩都在畫畫。 'All the boys were drawing.'	The difference and usage of the quantifiers "all", "some", and "every"	□1 □2 □3 □4 □5	
S07	Double-object Construction	雙賓句	小明送給老師一束花。 'Siu Ming gave the teacher a bouquet of flowers.'	The word order of double-object constructions	□1 □2 □3 □4 □5	
S08	Locative Existential	處所存在句	操場上站著一個男孩。 'There is a boy standing in the playground.'	The meaning of sentences with inanimate (location) subject and their difference with declarative sentences	□1 □2 □3 □4 □5	

S09	Control	控制句 兼語句？	小明要姐姐講故事。 ‘Siu Ming asked his sister to tell a story.’	The meaning of control sentences and their grammatical characteristics	□1 □2 □3 □4 □5	
S10	Cleft Sentences	分裂句	小明是後天參加比賽的。 ‘Siu Ming will participate in a competition the day after tomorrow.’	The meaning and grammatical characteristics of split sentences	□1 □2 □3 □4 □5	
S11	Question & Modal	疑問句	媽媽怎麼會去學校？ ‘How did mom go to school?’	The meaning of interrogative with modal verbs	□1 □2 □3 □4 □5	
S12	Morpheme Distinction	結構助詞	小明笑得很開心。 ‘Siu Ming smiled happily.’	The difference usage between the three structural particles	□1 □2 □3 □4 □5	
S13	Negation	否定詞	小明昨天沒有參加比賽。 ‘Siu Ming did not participate in the competition yesterday.’	The difference between the usage of the negators "no" and "not"	□1 □2 □3 □4 □5	
S14	Preposition	介詞	小明向公園跑去。 ‘Siu Ming is running towards the park.’	The distinction and usage of the prepositions "to", "follow", "from", "toward", and "at"	□1 □2 □3 □4 □5	
S15	Localizer	方位詞	小明坐在沙發上。 ‘Siu Ming is sitting on the sofa.’	The grammatical characteristics of locatives “up” and “inside”	□1 □2 □3 □4 □5	
S16	Aspect	體貌詞	小明喝了一杯水。 ‘Siu Ming has drunk a glass of water.’	The difference between the physical words. e.g. "in" and "in"	□1 □2 □3 □4 □5	
S17	Question words	疑問詞	小明什麼時候參加比賽？ ‘When does Siu Ming participate in the competition?’	The usage of different wh-words, such as "who", "what", or "why”, or “when” and where”	□1 □2 □3 □4 □5	
S18	Question Particles	疑問語氣詞	小明要不要參加比賽呢？ ‘Does Siu Ming want to participate in the competition?’	The grammatical characteristics of different question particles	□1 □2 □3 □4 □5	

1. According to the Chinese reading and writing development of primary school students, do you think the selected 18 Chinese grammatical items representativeness is adequate?

□ very poor representativeness □ poorly representative □ fairly representativeness

□ high representativeness □ very highly representativeness

2. Which grammatical items do you think need to be removed?

____

____

3. Is there any other grammatical items to be added? Please make suggestion.

____

____

4. If you have any other comments, please fill in below:

<u>Relevance of the Questions</u>

"Relevance" refers to whether the tests items used for the assessment can properly reflect the targeted grammatical knowledge. In the case of CGA, it refers to the students' Chinese grammatical knowledge. The three aspects that “relevance” concerns are: (1) the objectives of the assessment, (2) the theory behind the assessment, and (3) the elements incorporated in the items (including the content, question types, choices of answers, and answering methods, etc.) that can target on the specific Chinese grammatical knowledge. Chinese grammatical competence is composed of different facets of grammatical knowledge. For example, in the sentence, the word *了 (liu5)* in the sentence「妹妹吃了飯」(sister has eaten already) is an aspect marker, it modifies the verb and signifies the meaning of completion. It is an important grammatical knowledge in Chinese (Huang, Li, & Li, 2009). It is also a topic frequently studied in language acquisition research. Therefore, the items that test for the function of *了(liu5)* is highly relevant. If the same sentence「妹妹吃了飯」(sister has eaten already) is used as the assessment item, but the focus is mainly on the meaning of the verb *吃飯* (eating rice), the relationship between the item and the assessment objective is weak. The relevance of the item is thus considered very low (please refer to Figure 3).

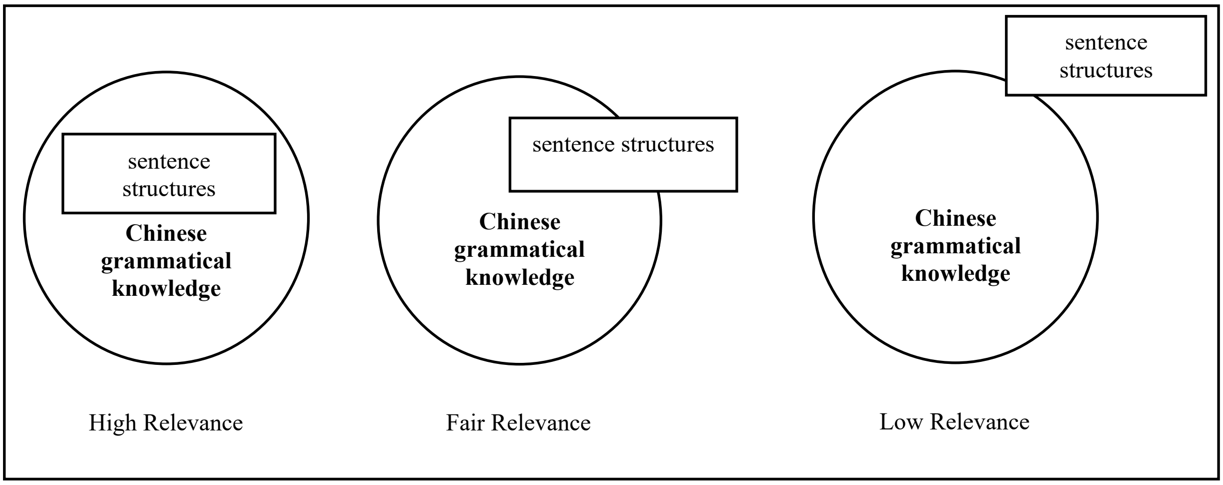


Figure 3: The relevance of the items

In the following part, we will list out all the 172 items according to their different task types. Please review them to see if they are having high relevance to the Chinese grammatical knowledge of primary school students:

1 = very poor relevance 2 = poor relevance 3 = fair relevance

4 = good relevance 5 = very good relevance

In addition, please also check the appropriateness of the design of each test items, that is, whether the question can effectively assess the knowledge of the target grammar:

1 = very poor appropriateness 2 = poor appropriateness 3 = fair appropriateness.

4 = high appropriateness 5 = very high appropriateness

(The items will be reviewed one by one following the four different task types)

Grammatical knowledge	Item	All choices (The correct answer is highlighted)	Relevance to Chinese grammatical knowledge	Whether the design of the items is appropriate#	Other comments about the test items (if any).
			□1 □2 □3 □4 □5	□1 □2 □3 □4 □5	
			□1 □2 □3 □4 □5	□1 □2 □3 □4 □5	
			□1 □2 □3 □4 □5	□1 □2 □3 □4 □5	
			□1 □2 □3 □4 □5	□1 □2 □3 □4 □5	
			□1 □2 □3 □4 □5	□1 □2 □3 □4 □5	
			□1 □2 □3 □4 □5	□1 □2 □3 □4 □5	

Part III

Overall, what is your opinion on the CGA? We hope that you can give us some opinions to help us improve the design of the entire assessment and the items used for assessment.

1) How appropriate is the title "Chinese Grammatical Assessment (中文語法評估)"?

□ very poor appropriateness □ poor appropriateness □ fair appropriateness

□ high appropriateness □ very high appropriateness

2) How appropriate are the overall operations of the assessment?

□ very poor appropriateness □ poor appropriateness □ fair appropriateness

□ high appropriateness □ very high appropriateness

3) Are the selected 18 grammatical categories of CGA having good representativeness in assessing primary school students' grammatical development in written Chinese?

□ very poor appropriateness □ poor appropriateness □ fair appropriateness

□ high appropriateness □ very high appropriateness

4) Are test items of CGA suitable for testing Chinese grammatical knowledge of deaf or hard-of-hearing children?

□ very poor appropriateness □ poor appropriateness □ fair appropriateness

□ high appropriateness □ very high appropriateness

5) Overall, do you have any other suggestions for the Chinese Grammar Assessment? Please fill in below.

Reference:

Huang, C.-T. J., Li, Y.-H. A., & Li, Y. (2009). *The Syntax of Chinese*. Cambridge University Press.

Appendix C: Items selection for CGA-A and CGA-B after psychometric review

Grammatical categories	Item Codes	Logit	CGA-A	CGA-B
S01	***ba*-constructions**			
	babvGJ03	52.21	A	
	babvGJ04	54.89		B
	bacoGJ01	43.50	A	
	bacoGJ02	37.69		B
	baxxTV03	51.17		B
	baxxTV04	50.67	A	
S02	**Passives**			
	beixTV02	45.30		B
	beixTV04	43.96	A	
	bpspPM02	35.93	A	
	bpspPM03	44.42		B
S03	**Binding**			
	bncrPS01	40.97		B
	bncrPS04	44.02	A	
	bnpnPS01	44.75	A	
	bnpnPS03	43.08		B
	bnpnPS07	46.36	A	
	bnpnPS08	50.16		B
	bnrfPS02	42.09		B
	bnrfPS04	41.51	A	
	bnrfPS05	40.55	A	
	bnrfPS07	40.05		B
S04	**Relative clause**			
	rcooPS01	45.08	A	
	rcooPS02	45.25		B
	rcosPS02	49.65	A	
	rcosPS03	49.60		B
	rcsoPS01	50.02	A	
	rcsoPS04	47.68		B
	rcssPS01	46.97		B
	rcssPS03	47.13	A	

Grammatical categories	Item Codes	Logit	CGA-A	CGA-B
S05	**Comparatives**			
	cnbbPM01	43.96	A	
	cnbbPM02	44.86		B
	cnbbPM03	63.86		B
	cnbbPM04	64.19	A	
	cnbbPM05	47.03	A	
	cnbbPM06	47.53		B
	cnmyPM03	43.61	A	
	cnmyPM04	43.32		B
	compPS03	41.44	A	
	compPS04	40.48		B
S06	**Quantification**			
	nqnqPS03	47.58	A	
	nqnqPS04	47.77		B
	nqqnPS02	42.03	A	
	nqqnPS03	40.48		B
	qualTV01	38.61		B
	qualTV04	39.23	A	
	quevTV02	40.34		B
	quevTV04	38.44	A	
S07	**Double-object construction**			
	docxWR02	46.31	A	
	docxWR03	44.25		B
S08	**Locative existential**			
	locaWR01	54.84	A	
	locaWR02	55.20		B
	lociWR01	47.48	A	
	lociWR02	49.60		B
S09	**Control**			
	ctocPS03	39.84	A	
	ctocPS04	41.31		B

Grammatical categories	**Item Codes**	**Logit**	**CGA-A**	**CGA-B**
S10	**Cleft sentences**			
	clseSC02	48.22	A	
	clseSC03	48.79		B
S11	**Question**			
	qmmaSC01	51.63		B
	qmmaSC03	50.58	A	
	qmreSC01	53.60		B
	qmreSC04	48.31	A	
S12	**Morpheme distinction**			
	mdeiFB02	43.85	A	
	mdeiFB04	43.61		B
	mdexFB03	46.05		B
	mdexFB04	45.14	A	
	mdixFB02	48.65		B
	mdixFB04	41.64	A	
S13	**Negation**			
	negbFB03	41.37	A	
	negbFB04	48.99		B
	negmFB02	38.77		B
	negmFB03	40.69	A	
S14	**Preposition**			
	precFB03	51.63	A	
	precFB04	48.31		B
	predFB01	46.57	A	
	predFB02	44.19		B
	pregFB01	42.03		B
	pregFB02	41.44	A	
	prexFB03	56.36		B
	prexFB04	55.69	A	
	prezFB03	51.08		B
	prezFB04	50.16	A	

Grammatical categories	Item Codes	Logit	CGA-A	CGA-B
S15	**Localizer**			
	loloGJ01	46.97	A	
	loloGJ04	44.08		B
	lonlGJ02	66.82	A	
	lonlGJ03	66.82		B
S16	**Aspect**			
	aspfTV07	37.95		B
	aspfTV08	44.70	A	
	aspgTV03	39.16		B
	aspgTV04	44.53	A	
S17	**Question words**			
	qwadFB01	48.84		B
	qwadFB04	45.14	A	
	qwarFB02	44.48	A	
	qwarFB04	43.44		B
S18	**Question particles**			
	qpmaGJ03	54.13		B
	qpmaGJ04	54.44	A	
	qpneGJ01	45.99	A	
	qpneGJ02	44.08		B

Appendix D: Percentile Ranks Calculated According to DHH Students' Reading Scores

Below are the percentile ranks developed based on students' reading scores in their final school examinations.

Grade Levels: P1 (N=17)

Raw Scores	Frequency	Percent	Valid Percent	Cumulative Percent
8	1	5.9	5.9	5.9
15	1	5.9	5.9	11.8
16	2	11.8	11.8	23.5
20	3	17.6	17.6	41.2
22	1	5.9	5.9	47.1
24	1	5.9	5.9	52.9
25	1	5.9	5.9	58.8
26	2	11.8	11.8	70.6
27	1	5.9	5.9	76.5
28	2	11.8	11.8	88.2
30	2	11.8	11.8	100.0
Total	17	100.0	100.0	

Grade Levels: P2 (N=19)

Raw Scores	Frequency	Percent	Valid Percent	Cumulative Percent
12	2	10.5	10.5	10.5
14	1	5.3	5.3	15.8
16	2	10.5	10.5	26.3
17	1	5.3	5.3	31.6
18	1	5.3	5.3	36.8
20	1	5.3	5.3	42.1
22	1	5.3	5.3	47.4
23	2	10.5	10.5	57.9
24	1	5.3	5.3	63.2
25	1	5.3	5.3	68.4
26	1	5.3	5.3	73.7
27	1	5.3	5.3	78.9
28	1	5.3	5.3	84.2
29	1	5.3	5.3	89.5
30	2	10.5	10.5	100.0
Total	19	100.0	100.0	

Grade Levels: P3 (N=23)

Raw Scores	Frequency	Percent	Valid Percent	Cumulative Percent
6	1	4.3	4.5	4.5
7	1	4.3	4.5	9.1
10	1	4.3	4.5	13.6
12	1	4.3	4.5	18.2
16	1	4.3	4.5	22.7
17	1	4.3	4.5	27.3
18	1	4.3	4.5	31.8
19	2	8.7	9.1	40.9
20	1	4.3	4.5	45.5
22	3	13.0	13.6	59.1
25	2	8.7	9.1	68.2
26	2	8.7	9.1	77.3
27	1	4.3	4.5	81.8
28	3	13.0	13.6	95.5
29	1	4.3	4.5	100.0
Total	22	95.7	100.0	
System	1	4.3		
	23	100.0		

Grade Levels: P4 (N=24)

Raw Scores	Frequency	Percent	Valid Percent	Cumulative Percent
7	1	4.0	4.2	4.2
8	1	4.0	4.2	8.3
10	1	4.0	4.2	12.5
14	1	4.0	4.2	16.7
17	2	8.0	8.3	25.0
18	1	4.0	4.2	29.2
19	1	4.0	4.2	33.3
20	1	4.0	4.2	37.5
21	3	12.0	12.5	50.0
22	1	4.0	4.2	54.2
23	1	4.0	4.2	58.3
24	4	16.0	16.7	75.0
26	2	8.0	8.3	83.3
27	4	16.0	16.7	100.0
Total	24	96.0	100.0	
System	1	4.0		
	25	100.0		

Grade Levels: P5 (N=23)

Raw Scores	Frequency	Percent	Valid Percent	Cumulative Percent
18	1	4.3	4.3	4.3
19	1	4.3	4.3	8.7
22	1	4.3	4.3	13.0
23	1	4.3	4.3	17.4
24	2	8.7	8.7	26.1
26	4	17.4	17.4	43.5
28	5	21.7	21.7	65.2
30	8	34.8	34.8	100.0
Total	23	100.0	100.0	

Grade Levels: P6 (N=22)

Raw Scores	Frequency	Percent	Valid Percent	Cumulative Percent
6	1	4.5	4.5	4.5
14	1	4.5	4.5	9.1
16	2	9.1	9.1	18.2
18	2	9.1	9.1	27.3
19	1	4.5	4.5	31.8
20	1	4.5	4.5	36.4
23	2	9.1	9.1	45.5
24	2	9.1	9.1	54.5
26	3	13.6	13.6	68.2
27	2	9.1	9.1	77.3
28	1	4.5	4.5	81.8
29	1	4.5	4.5	86.4
30	3	13.6	13.6	100.0
Total	22	100.0	100.0	

Appendix E: Percentile Ranks Calculated According to DHH Students' Writing Scores

Below are the percentile ranks developed based on students' writing scores in their final school examinations.

Grade Level: P1 (N=17)

Raw Scores	Frequency	Percent	Valid Percent	Cumulative Percent
23	1	5.9	5.9	5.9
30	1	5.9	5.9	11.8
31	1	5.9	5.9	17.6
33	1	5.9	5.9	23.5
35	1	5.9	5.9	29.4
36	1	5.9	5.9	35.3
37	1	5.9	5.9	41.2
38	1	5.9	5.9	47.1
39	1	5.9	5.9	52.9
40	1	5.9	5.9	58.8
43	1	5.9	5.9	64.7
44	2	11.8	11.8	76.5
45	2	11.8	11.8	88.2
46	1	5.9	5.9	94.1
47	1	5.9	5.9	100.0
Total	17	100.0	100.0	

Grade Level: P2 (N=19)

Raw Scores	Frequency	Percent	Valid Percent	Cumulative Percent
27	1	5.3	5.3	5.3
31	1	5.3	5.3	10.5
32	1	5.3	5.3	15.8
37	1	5.3	5.3	21.1
38	2	10.5	10.5	31.6
39	1	5.3	5.3	36.8
40	1	5.3	5.3	42.1
41	1	5.3	5.3	47.4
43	3	15.8	15.8	63.2
44	1	5.3	5.3	68.4
45	1	5.3	5.3	73.7
46	2	10.5	10.5	84.2
47	2	10.5	10.5	94.7
48	1	5.3	5.3	100.0
Total	19	100.0	100.0	

Grade Level: P3 (N=22)

Raw Scores	Frequency	Percent	Valid Percent	Cumulative Percent
17	1	4.3	4.5	4.5
31	2	8.7	9.1	13.6
37	1	4.3	4.5	18.2
38	2	8.7	9.1	27.3
39	3	13.0	13.6	40.9
40	1	4.3	4.5	45.5
41	4	17.4	18.2	63.6
42	2	8.7	9.1	72.7
44	2	8.7	9.1	81.8
45	1	4.3	4.5	86.4
46	1	4.3	4.5	90.9
47	2	8.7	9.1	100.0
Total	22	95.7	100.0	
System	1	4.3		
	23	100.0		

Grade Level: P4 (N=24)

Raw Scores	Frequency	Percent	Valid Percent	Cumulative Percent
16	1	4.0	4.2	4.2
17	1	4.0	4.2	8.3
21	1	4.0	4.2	12.5
25	1	4.0	4.2	16.7
27	1	4.0	4.2	20.8
28	1	4.0	4.2	25.0
29	3	12.0	12.5	37.5
32	1	4.0	4.2	41.7
34	2	8.0	8.3	50.0
35	1	4.0	4.2	54.2
36	1	4.0	4.2	58.3
37	1	4.0	4.2	62.5
38	3	12.0	12.5	75.0
39	1	4.0	4.2	79.2
40	1	4.0	4.2	83.3
44	3	12.0	12.5	95.8
49	1	4.0	4.2	100.0
Total	24	96.0	100.0	
System	1	4.0		
	25	100.0		

Grade Level: P5 (N=23)

Raw Scores	Frequency	Percent	Valid Percent	Cumulative Percent
29	1	4.3	4.3	4.3
33	1	4.3	4.3	8.7
36	1	4.3	4.3	13.0
37	1	4.3	4.3	17.4
39	2	8.7	8.7	26.1
40	1	4.3	4.3	30.4
41	3	13.0	13.0	43.5
42	1	4.3	4.3	47.8
43	3	13.0	13.0	60.9
44	2	8.7	8.7	69.6
45	1	4.3	4.3	73.9
46	1	4.3	4.3	78.3
47	3	13.0	13.0	91.3
48	1	4.3	4.3	95.7
49	1	4.3	4.3	100.0
Total	23	100.0	100.0	

Grade Level: P6 (N=22)

Raw Scores	Frequency	Percent	Valid Percent	Cumulative Percent
9	1	4.5	4.5	4.5
20	1	4.5	4.5	9.1
23	2	9.1	9.1	18.2
24	1	4.5	4.5	22.7
26	1	4.5	4.5	27.3
27	1	4.5	4.5	31.8
29	1	4.5	4.5	36.4
35	1	4.5	4.5	40.9
36	2	9.1	9.1	50.0
37	2	9.1	9.1	59.1
38	1	4.5	4.5	63.6
41	2	9.1	9.1	72.7
42	2	9.1	9.1	81.8
44	1	4.5	4.5	86.4
45	1	4.5	4.5	90.9
47	1	4.5	4.5	95.5
50	1	4.5	4.5	100.0
Total	22	100.0	100.0	