Evaluating an Expressive Language Screening Tool for Five-year-old Cantonese-Speaking Children

Lam Yin Ling

A dissertation submitted in partial fulfillment of the requirements for the Bachelor of Science (Speech and Hearing Sciences), The University of Hong Kong, June 30, 2010.

Abstract

This study examined the screening accuracy of the Cantonese Early Identification Test for Pre-primary Children (CEIT) in identifying Cantonese-speaking preschool children with language impairment (LI). Nineteen children, aged between 5;01 and 5;06, from 2 kindergarten-cum-nurseries and an Early Education Training Centre, were administered the CEIT and the HKCOLAS's Cantonese Grammar subtest. The CEIT was found to have high sensitivity (100%), fair specificity (77%), and modest LR+ (4.35) and LR- (0.23) values if the "gold" standard for the child's language status was determined by the Cantonese Grammar subtest score. This finding provides further evidence that the CEIT is a potentially useful language screening tool for preschool children in Hong Kong. Further research should examine its screening accuracy with a larger and broader age group of children, and with the use of current performance in objective measures as the "gold" standard of language status.

Evaluating an Expressive Language Screening Tool for Five-year-old Cantonese-Speaking Children

Early identification and referral of language-impaired cases to intervention programs are important in minimizing the negative impacts of language impairment on the children's communication, psychosocial, personality and literacy development (Lin, 1994). Screening is an important initial step towards early intervention. Screening, by definition, is a brief initial test used to determine whether further diagnostic testing is necessary and possibly to guide the selection of other tests to be administered (Klee, 2008). These goals are best met with screening tools that demonstrate good psychometric properties and screening accuracy.

Current Language Screening Tool in Hong Kong

Before 2008, there was only one standardized preschool language screening tool for Hong Kong children, the Developmental Language Screening Scale (DLSS) (Lee, Luk, Yu & Bacon-Shone, 1985). The screening tool is in parent-report format and it assesses comprehension and expression competencies of 3-year-old children. However, DLSS was criticized for a number of reasons in Fong's study (2007). First, DLSS utilized age-equivalent scores, instead of standard scores, to define language impairment. This utilization of "discrepancy-based criteria" (Wong et al., 1992) was criticized to be inappropriate for deciding whether a child has a significant deficit since the score did not include any measure of normal variation (Paul, 2007). Besides, DLSS has been developed for 25 years in which

outdated norms are suspected. Lastly, poor screening accuracy data (sensitivity of 75% and specificity of 50%) were reported in the small scale study by Fong (2007). Therefore, there is an immediate need to develop another locally applicable screening tool with good evidence for accurately identifying children with language impairment.

Cantonese Oral Language Deficiency Early Identification Test for Pre-primary Children

The test under investigation, the Cantonese Oral Language Deficiency Early Identification Test for Pre-primary Children (CEIT) (PLKDSTT, 2008), is the second locally published standardized language screening test. This test uniquely involves kindergarten teachers and special child care workers as test administrators. Its purpose is to help teachers to detect preschool children aged from three to six at risk of language impairment, so that early referral can be made for a more in-depth evaluation and intervention by a speech therapist. According to the test manual, the item contents were mainly selected from the curriculum adopted in Po Leung Kuk kindergartens with additional reference to two local studies on oral language development in preschool children (Fletcher, Leung, Stokes & Weizman, 2000; Tse, 2006). The test contains 11 sections, which measure a wide range of expressive language knowledge, including vocabulary (nouns, verbs, adjectives, super-ordinates and classifiers), sentence structures (active sentence, passive sentence, comparative sentence, double object construction and relative clauses) and narrative. There are between two and six items in each section and a total of 49 items altogether. The test

adopts a dichotomized scoring method (i.e. correct and incorrect responses) and children scoring below -1 SD for age are considered having failed the screening and should be referred for follow-up evaluation.

A random-cluster sample of 286 typically developing preschool children in ten Po Leung Kuk kindergarten cum nurseries (153 males and 133 females) contributed to the normative data in CEIT. These children aged between three to six years old and they fell in groups separated by six-month intervals. In each interval, there were variations in sample size, ranging from 26 (3;01-3;06) to 65 (3;07-4;00). The normative data were then summarized and analyzed using curvilinear regression to obtain the mean raw score and standard deviation for each month interval. The concurrent validity of the test was examined using the expressive quotient of the Cantonese version (RDLS-C; Hong Kong Society for Child Health and Development, 1987) of the Reynell Developmental Language Scale (Reynell & Huntley, 1995) and a high positive correlation of 0.76 was reported (PLKDSTT, 2008). Both special child care workers and speech therapists were involved in the normative sampling process. Although the teachers' scoring was mentioned to be generally higher than that of the speech therapists, no specific inter-rater reliability figures were reported.

Evidence-Based Practice

In recent years, the American Psychological Association advocates evidence-based practice (EBP) in decisions regarding assessment or intervention. Although screening results

alone cannot determine the existence of language impairments, Dollaghan (2007) considered screening as “a kind of classification tool” (p.81) which should be evaluated against an evidence-based framework as well. Since screening is an initial step for the whole evaluation, its results should be as accurate as possible to minimize the expenses associated with both over-referrals and under-detection. Screening errors would either prevent children with language impairment from receiving follow-up evaluation and early intervention (Anderson et al., 2003), or create needless anxiety in the parents and the children who actually did not have language impairment and result in wastage of professional time and resources.

There are two complementary ways to evaluate the test adequacy as suggested by Klee, Wong, Stokes, Fletcher and Leonard (2009). The first one is a psychometric review of the test. McCauley and Swisher (1984) were the first to apply this approach to evaluate test adequacy using 10 well-established psychometric criteria. The psychometric criteria include areas of the representativeness of normative sample, test procedures, data normalization as well as the test’s reliability and validity (Plante & Vance, 1994).

The second approach is an evidence-based one which mainly evaluates the diagnostic accuracy and clinical usefulness of the test (Klee et al., 2009). Critical evidence for screening accuracy comes from a comparison between an *index measure* (the screening test being evaluated) and a *reference standard*, which refers to a gold standard in establishing the presence of the target condition (Bossuyt et al., 2003). Accuracy can be measured in terms of

the tool's ability to adequately identify children with language impairment as being language impaired (sensitivity) and to identify typically-developed children as being normal (specificity). Sensitivity and specificity values of 80% are generally accepted as the minimum level for discriminate accuracy (Plante & Vance, 1994). In addition, Dollaghan (2007) recommended using positive (LR+) and negative likelihood ratios (LR-) as accuracy measures since they are less susceptible to variations in the prevalence of the diagnostic condition in the sample. Likelihood ratios refer to the probability ratios of obtaining a specific test result from the population who do have the disorder to that from the population who do not have the disorder. McAlister, Straus & Sackett (1999) suggested that a LR+ larger than 10 and a LR- less than 0.1 with reasonable confidence intervals are the minimal levels acceptance for an index test. To effectively evaluate the diagnostic adequacy of any screening or diagnostic tests from an evidence-based approach, Whiting, Rutjes, Reitsma, Bossuyt and Kleijnen (2003) described a 14-point checklist QUADAS (Quality Assessment of Diagnostic Accuracy Studies) (Appendix B) to help test users appraise the test's methodological quality.

Regarding the screening accuracy of CEIT, no specific figures on sensitivity, specificity, or likelihood ratios were reported in the manual (PLKDSTT, 2008). According to the 14-item QUADAS checklist (Whiting et al., 2003), the original CEIT study fell short in two areas: the usage of inappropriate and non-concurrent reference standard and the lack of clear blinding procedures. Although it was reported that the test was administered to 36 children in the

"Integrated Programme" (IP) with mild disabilities and 30 of them were screened positive, the use of IP status as the "reference standard" may be misleading since some students enrolled in the IP are with physical handicap, hearing impairment and visual impairment, in which language impairments are not necessarily implied. Moreover, the administration of both the "reference standard" (enrollment criteria for IP) and the "index test" (CEIT) may not be concurrent for all 36 children so that there was a possibility that a change in the language status of some of these children may result, especially when they were receiving speech therapy. Lastly, since no blinding process in the validation process was reported, it is plausible that the information reported in the CEIT regarding its screening accuracy could have been contaminated by subjective bias of the test examiners. Lack of such data for CEIT prevents clinicians from evaluating the potential use of this test in an evidence-based manner.

With such considerations, the present study evaluated the test adequacy of the Cantonese Oral Language Deficiency Early Identification Test for Pre-primary Children (PLKDSTT, 2008) using both psychometric criteria and diagnostic accuracy criteria. Data of sensitivity, specificity and likelihood ratios were presented to enable evidence-based practice in the future use of this screening tool and to inform future research on language screening. The concurrent validity of the CEIT was also examined by providing correlations between scores on the CEIT and those in a more up-to-date standardized language test. Besides, since different word classes and syntactic structures were tested in CEIT, the present study aimed to

analyze the error patterns observed from both groups of typically-developed children (TD) and children with language impairments (LI) in each tested area, to identify potential diagnostic language markers for late-preschool children with language impairment.

In CEIT, there are six normative subgroups between 3;00 and 6;00 in every six-month interval. Due to time limitation in the present study, its screening accuracy was examined only in one subgroup. The choice was between group 5 (5;01-5;06) and group 6 (5;07-6:00), since the normative samples of HKCOLAS, the reference test in this study (to be discussed later) were 5;00 and above. Since a larger sample will give more reliable and precise results than measures derived from smaller samples (Plante & Vance, 1994), the group with a larger normative sample size, i.e. group 5 (5;01-5;06, n = 54) was chosen over group 6 (5;07-6;00, n = 27), as the target group in the present study. This study aimed to examine: (1) The test adequacy of the CEIT in identifying 5;01-5;06 children with language impairment, (2) The concurrent validity of the CEIT in identifying 5;01-5;06 children with language impairment, and (3) The performance and error patterns of TD and LI groups in each test area of CEIT.

Method

Participants

All children aged between 5;01 and 5;06 in two kindergarten-cum-nurseries (nurseries) and one Early Education Training Centre (EETC) were invited to participate in the study. The EETC admits only 2;00-6;00 children with developmental delay or mental retardation for

non-school basis training. The two kindergarten-cum-nurseries invited were both Integrated Child Care Centers (ICCC) with admission of both normally-developing and integrated program (IP) children. Parents of the participating children completed a case history form (Appendix). Children who were diagnosed with pure speech impairment and children who were reported as bilingual were excluded. Eventually 19 children participated in the study.

Measures and Procedures

The presented study was conducted in two stages on two separate days either in the children's kindergarten-cum-nurseries or in their home (for children recruited from the EETC). The location arrangement helped ensure blinding of the clinical status of the children.

Stage I Screening

Stage I involves the administration of the index screening test, the Cantonese Oral Language Deficiency Early Identification Test for Pre-primary Children (CEIT) (PLKDSTT, 2008). Given that it was not possible to recruit a sufficient number of teachers for the study, three speech therapy undergraduate students were invited as the test administrators. This modification should not affect the validity of the screening results since speech therapists were actively involved in the test design, data collection, normative sampling and teacher training in the original validation process (PLKDSTT, 2008). In stage I, the test administrators were totally blind to the children's clinical status. They were trained in group for the administration of CEIT by the principle investigator before data collection to ensure

comparable training background and to minimize examiner effects. Screening sessions lasted for 15-20 minutes. Children scoring less than one standard deviation below mean were screened as at-risk of language impairment as defined in the test manual (PLKDSTT, 2008).

Stage II Reference standard

There were two reference standards employed in present study, including the current clinical status of the children and the test of Hong Kong Cantonese grammar in the Hong Kong Cantonese Oral Language Assessment Scale (HKCOLAS; T'sou et al., 2006). The subject's current clinical status diagnosed by speech therapists was the primary reference standard in the present study since Lund and Duchan (1993) suggested that an objective gold standard should incorporate test administrators' clinical impression in diagnosis to reduce the effect of performance factors in one single test. Children who were diagnosed as having language impairment and were receiving language intervention were classified positive cases according to this reference standard. However, since the criterion or tests used in their previous clinical diagnosis were unknown and the time of the latest language diagnosis was variable for each of the children, the Cantonese grammar subtest in HKCOLAS was used as another reference standard in this study to provide up-to-date and consistent information about the child's current language ability. The correlation between the scores obtained in CEIT and those in Cantonese grammar test in the HKCOLAS help provide additional information about the construct and concurrent validity of the CEIT.

HKCOLAS was a locally-designed standardized test published in 2006 for assessing Cantonese-speaking children between 5 to 12 years old. Its test norms were based on 1120 children and high reliability and validity were reported (T'sou et al., 2006). HKCOLAS provides more up-to-date norms than RDLS-C (Hong Kong Society for Child Health and Development, 1987), which is another standardized language test for children under six. There are in total six core subtests in HKCOLAS, including test of Hong Kong Cantonese grammar, textual comprehension test, word definition test, lexical-semantic test, expressive nominal vocabulary test and narrative test (T'sou et al., 2006). Children obtaining standard scores of less than 1.25 SD below mean in two or more subtests in HKCOLAS are diagnosed as having language-impairment.

However, administration of the whole battery requires around 60-90 minutes and due to time limitation, only the Cantonese grammar subtest was used as the reference standard test in the present study. This particular subtest was chosen because it taps the child's knowledge in a wide range of word classes and Cantonese grammar, which are similar to the test construct of CEIT. Also, the large numbers of test items as well as their diverse difficulty levels enable this test to capture a spectral performance for both TD and LI groups with minimal interference from floor or ceiling effects, which outperform other shorter subtests in HKCOLAS. Lastly, this subtest utilizes an easy and objective scoring method which is likely to reduce the subjectivity and inter-rater differences during scoring. An ongoing research

project on 5-6 year old children with SLI (A.M-Y. Wong, personal communication, February 12, 2010) suggested that in a sample of 37 children who met the criteria for SLI (failed at least 2 subtests in the HKCOLAS and met other conventional criteria for SLI), 28 received a standard score lower than -1.25 SD in the Cantonese grammar subtest. This suggested a good potential of this subtest in identifying five to six years old children with SLI. In this study, children who scored lower than -1.25 SD on the Cantonese grammar subtest would be considered language impaired.

In stage II, the principle investigator was the test administrator who was blind to the subject's screening results in stage I to minimize subjectivity and bias. To ensure proper administration of the HKCOLAS's Cantonese grammar subtest, an experienced speech therapist (the supervisor of this study) was invited to observe and evaluate the test's administrator's running of the reference test on a seven-year-old child in a pilot session, after which the test administrator's interpretation of test results were also discussed. This stage lasted for 30-40 minutes.

Inter-rater Reliability of Index Test

Inter-rater reliability of index test scores was evaluated by comparing scores from two test administrators on two randomly selected children. Percentage agreement (dividing the number of agreement by the total number of items) was used to evaluate inter-rater reliability due to the mutually exclusive categories (either correct or incorrect) classified by the

responses. The percentage agreement of the test administrator's rating on the CEIT was 96%. This high percentage agreement indicated good inter-rater reliability.

Results

Nineteen children between the age of 61 and 66 months (mean = 62.2 months) completed the study. Using their current clinical status as the reference standard, 10 of them (three boys, seven girls) were considered typically-developing (TD) and the remaining nine (eight boys, one girl) as having language impairments (LI). After applying a cut-off point of one standard deviation below mean in the screening test CEIT, eight of the ten (80%) TD cases were screened negatives while seven of the nine LI (77.8%) cases were screened positives. If, however, we used -1.25 SD in HKCOLAS's Cantonese grammar subtest as the reference standard, 13 of the 19 children (seven boys and six girls) would be considered TD and the other six (four boys and two girls) as being LI, resulting in some improvement in CEIT's screening accuracy. Ten of the 13 (77%) TD cases being screened negatives while all LI (100%) cases being screened positives in CEIT.

Discriminant Analysis

The screening accuracy of CEIT can be examined with the use of "Diagnostic 2 x 2 table" shown in table 1 (Dollaghan, 2007; Klee, 2008). Several indices including sensitivity, specificity, positive likelihood ratio (LR+) and negative likelihood ratio (LR-) can be computed using the above table. Sensitivity refers to the proportion of tested positive cases

Table 1. Diagnostic 2 x 2 table

		By reference standard (Current clinical status, Grammar subtest)	
		Tested Positive	**Tested Negative**
By index test (CEIT)	**Screened Positive**	True Positive (a)	False Positive (b)
	Screened Negative	False Negative (c)	True Negative (d)

being accurately identified by the screening test, i.e. (a/ (a + c)) while specificity refers to the proportion of tested negative cases being accurately identified, i.e. (d / (b + d)). LR+ refers to the proportion of positive screening results came from the children with language impairment, i.e. (sensitivity / (1-specificity)) whereas LR- refers to the proportion of negative screening results actually came from the children with typical language development, i.e. ((1-sensitivity) / specificity). Table 2 shows the 95% confidence interval (CI) for the sensitivity, specificity, LR+ and LR- for the CEIT when using either clinical status or Cantonese grammar subtest as reference standard.

The difference in performance between the screened-positive and the screened-negative groups in the CEIT and the Cantonese grammar subtest were further examined using Mann-Whitney U test with statistical significance set at $p < 0.05$. Table 3 gives the descriptive statistics of the two groups' CEIT and Cantonese grammar subtest scores. Results showed that the subgroups differed significantly in both CEIT ($U = 0.5$, $n_1=9$, $n_2=10$, $p< 0.05$)

Table 2. Confidence intervals (CIs) for accuracy metrics under different reference standards

	Clinical status		Cantonese grammar subtest	
Accuracy metric	Value	95% CI	Value	95% CI
Sensitivity	0.778	0.453 to 0.937	1.0	*
Specificity	0.8	0.49 to 0.943	0.77	*
Positive likelihood ratio (LR+)	3.889	1.073 to 14.097	4.35	*
Negative likelihood ratio (LR-)	0.278	0.079 to 0.98	0.23	*

Note: CI = Confidence interval

* in 95% CI indicate the uncertainty and a lack of evidence on precision when either sensitivity or specificity is perfect (1.0 or 100%)

and Cantonese grammar scores ($U = 3.5$, $n_1=9$, $n_2=10$, $p < 0.05$). Cliff's delta was used as a nonparametric effect size measure to estimate the degree of overlapping between two scores' distributions (Cliff, 1993). Bridging between Cohen's *d* value and Cliff's delta, a delta value of 0.147 is interpreted as a small effect, 0.33 is a medium effect, and 0.474 is a large effect (Romano, Kromrey, Coraggio & Skowronek, 2006). In the present study, the effect size for the group difference on CEIT and the Cantonese Grammar subtest was large.

Concurrent Validity of CEIT

Person Product-Moment Correlation analysis was run to examine the degree of association between children's scores from the CEIT and those from the Cantonese grammar

Table 3. Comparison of the screening group performance in the CEIT and reference test

	Screened +ve (n=9)		Screened -ve (n=10)			
Measure	**mean**	**SD**	**mean**	**SD**	**U**	**Cliff's *d***
CEIT	23.11	11.12	41.4	3.86	0.50*	1**
Cantonese grammar test	23	4.95	38.1	7.08	3.50*	0.92**

Note: CEIT = Cantonese Oral Language Deficiency Early Identification Test for Pre-primary Children. +ve = Positive; -ve = Negative.

*$p < 0.05$, two-tailed. **Large effect size based on Cliff's delta

subtest, with statistical significance set at $p < 0.05$. High positive correlations with statistical significance ($r(17) = 0.81$, $p < 0.05$) indicated a good concurrent validity of CEIT.

Group Difference in CEIT test areas

Recall that the CEIT includes 11 sections, covering different language forms and structures. Table 4 gives the descriptive statistics of the groups' performance in each section. For this analysis, the LI group was defined using the children's Cantonese Grammar subtest scores. It is because this criterion provides more up-to-date information about the child's current language ability. Given the unequal and small sample size, the Mann-Whitney U tests, were used to examine the group differences. In order to keep the experiment-wise error rate to a specified level ($p = .05$) and control the Type I error (the error of incorrectly declaring a difference), bonferroni adjustment was used a priori to multiple Mann-Whitney U tests. The

Table 4. Comparison of the LI and TD groups' performance in each test area in CEIT

	LI group (n=6)		TD group (n=13)			
Measure (no. of test items)	**mean**	**SD**	**mean**	**SD**	**U**	**Cliff's *d***
Nouns (6)	5.17	1.60	5.69	0.63	34.0	0.13
Verbs (5)	2.67	1.5	3.77	1.42	21.0	0.46#
Adjectives (5)	3.33	1.75	4.61	0.77	16.5	0.22^
Super-ordinates (4)	2.67	1.50	3.23	1.17	29.0	0.05
Classifiers (4)	1.33	1.51	2.85	1.57	16.5	0.27^
Active Sentences (2)	1	0.89	1.77	0.60	19.0	0.5**
Comparative Sentences (4)	0.83	1.17	2.31	1.18	13.0	0.67**
Passive Sentences (5)	1.33	1.51	4.23	1.42	5.5*	0.86**
Double Objects Sentences(5)	1.17	1.17	3.15	1.34	10.5	0.73**
Relative Clauses (6)	1.17	1.94	3.54	1.71	15.0	0.62**
Narration (3)	1.50	1.22	2.46	0.78	21.0	0.13

Note: TD = Typical-developed; LI = Language-impaired

*p<0.0045, two-tailed. **Large effect size # Medium effect size ^ Small effect size

new p-value was computed by dividing the overall significance level (p = .05) by the number of comparisons intended to make, i.e. 11 comparisons in the present study. The adjusted statistical significance for each comparison was $p < 0.0045$.

With this adjusted p value, the mean ranks of the TD and LI groups differed significantly only in "passive sentence" ($U = 5.5$, $n_1=6$, $n_2=13$, $p < 0.0045$). However, due to the abundant experiment-wise comparisons made in the present study, bonferroni adjustment was suggested to be too conservative in ascertaining a difference between groups and may result in large type II errors, i.e. declaring no effect or difference while in fact there is an effect. In this case, effect size would be more informative than significance testing. Based on Cliff's delta, all five sections examining sentence structures (active, passive, comparative, double object sentences and relative clauses) showed a large effect size while the section examining verbs showed a medium effect size.

Discussion

Adequacy of the CEIT as Language Screening Test

In this section, we will first discuss the CEIT's screening adequacy using the evidence-based approach, to be followed by an evaluation of its normative construction and psychometric properties. There was a large difference in CEIT's sensitivity when compared with two different reference standards, i.e. clinical status (78%) versus HKCOLAS's Cantonese grammar subtest (100%). This discrepancy may be explained by the time lag between the administrations of the reference standards and the screening test. When using the children's clinical status as reference, the time of being diagnosed as language-impaired was variable and not concurrent with the administration of the screening test. This may increase

the chance for the children to improve their language abilities with speech therapy so that they might perform as well as their age peers in the recently-administered Cantonese grammar subtest and CEIT. Nevertheless, the high sensitivity value (100%) in the concurrent reference standard, i.e. Cantonese grammar subtest, suggested that the CEIT is highly accurate in identifying 5-year-old children with language impairments.

CEIT's specificity was comparable when using either clinical status (80%) or Cantonese grammar subtest (77%) as the reference standard. Based on the benchmarks from Plante and Vance (1994), this indicated that CEIT has a borderline-to-acceptable accuracy in identifying 5-year-old children without language impairments. The slight tendency to over-refer cases may be due to the ceiling effects together with the elicitation controversy for the specific language structures "double objects sentences". In this section, the children were asked to describe all the objects and events in the given picture using one complete sentence.

E.g. Item 38. *Naam4 jan2 bei2 gau3 gwat1tau4 gau2zai2 sik6*

Man give (CL) bone doggie eat

The man gives the doggie a bone to eat.

Item 39. *Caak3zai2 coeng2 zo2 maa1mi4 go3 ngan4baau1*

Thief rob (ASP) mother (CL) wallet

The thief robs mother's wallet.

However, quite a few children produced alternative sentence structures, e.g. *Naam4 jan2 bei2*

gau2zai2 sik6 gwat1tau4 (inversion of the double objects) or *Caak3zai2 coeng2 zo2 maa1mi4 ge3 ngan4baau1* (substitution by one single object with possessive pronoun), which resulted in zero points. These structures however were syntactically correct and described the pictures equally unambiguously. Moreover, CEIT's maximum score is 49 and the mean of the normative group 5;01 to 5;06 age group was 43 with a standard deviation of six. Two children in the sample approached ceiling performance and scored 47 points. Given the very narrow range of scores for the older normative age group, a few ambiguous mistakes may drive the children out of the -1 SD cut-off and result in a false positive case.

To evaluate the CEIT's normative construction and psychometric adequacy, we applied the 10 psychometric criteria reported in Plante and Vance (1994) (Appendix C). CEIT met four of these criteria, including evidence of item analysis, report of mean and standard deviations in each normative age group, satisfactory concurrent validity and description of test procedures. In their study of the CEIT, PLKDSTT (2008) reported a significant positive correlation of 0.76 between the CEIT and the RDLS-C scores. The present study compared the CEIT and the HKCOLAS's Cantonese grammar subtest scores and found a statistically significant and strong correlation ($r = 0.81$) in the 5;01-5;06 age group, and hence further confirmed its validity.

However, the CEIT failed to meet the normative sample size criteria of at least one hundred children in each of the normative age groups. It also did not provide a clear

definition of the standardization population in terms of geometric region and socioeconomic backgrounds. Besides, the selection of samples from only Po Leung Kuk kindergarten-cum-nurseries during the norming process may cause sampling bias and hence reduce the representativeness of the sample to a larger population of preschool children in all local kindergartens and nurseries. Information about predictive validity and test-retest reliability was also not reported in the CEIT test manual. For inter-rater reliability, although no specific inter-rater reliability figures were provided, teachers' scoring was reported to be generally higher than that of the speech therapists. This information, nevertheless, is particularly important especially since kindergarten teachers was intended to be the primary users of this screening tool. There was also no mention of examiner qualifications and specialized training in the CEIT manual. This area will be further discussed in a later section.

Potential Diagnostic Language Markers for Five-year-old Children with LI

Different word classes and syntactic structures were tested in the CEIT and the results showed that 5;01-5;06 children with language impairments (LI) performed significantly poorer than typically-developed children (TD) in all syntactic structures examined, i.e. active sentences, passive sentences, comparative sentences, double object sentences and relative clauses. These findings reconfirmed the CEIT's construct validity and provided useful reference for the future development of language screening or diagnostic tests. Although several children in the LI group had other developmental problems, including Autistic

Spectrum Disorder and developmental delay, the results from this study may also give insights for future investigation of potential diagnostic language markers for late-preschool Cantonese-speaking children with SLI (CSLI). In the current literature, CSLI was found to have poorer aspect markers use (Fletcher, Leonard, Stokes & Wong, 2005), more frequent use of general instead of specific classifiers (Stokes & So, 1997) and more difficulty in passive sentence construction (Leonard, Wong, Deevy, Stokes & Fletcher, 2006) than their TD age-matched peers. There has been no published report on children's CSLI's use of comparative sentences, double object sentences and relative clauses construction. Given results from this study, it will be of interest to examine these structures in a larger sample of CSLI children.

Scott (2004) suggested that quantitative sentence-level measures such as mean length utterance (MLU) are not age-sensitive for children beyond 5 years of age. Rather, syntactic complexity measures and narrative performance are more reliable and valid indicators of language impairment during later language development. To (in press) also investigated the diagnostic accuracy of story telling skills, including macrostructure skills (i.e. story grammar) and four microstructure skills (i.e. syntactic complexity, semantic sore, referencing and connective use), in identifying CSLI children aged between 5 and 12 years old. High sensitivity value (86% to 94%), modest specificity (60% to 90%), LR+ (2.15 to 9.42) and LR- values (.07 to .34) were reported. However, in present study, the CEIT's "narration"

section did not show significant group difference and the effect size (d = 0.13) between the LI and TD group was negligible. This difference may be explained by the limited number and the relative ease of the test items. The three test items examined rather general information from the children's stories: naming of the story characters and objects, a clear and correct temporal sequence and the story theme. These items were relatively easy for the children in this age group, as most of the children in LI group were able to score at least 2 points, despite their limited and inappropriate syntactic construction or word use during story-telling.

Given that kindergarten teachers may not have extensive linguistic background and time or motivation to carry out complex linguistic analysis on the story sample, one way to improve the sensitivity of the narration section of the CEIT may be to score the children's use of specific vocabulary and connectives in a recognition format. The first measure assesses the child's ability to convey story information and his/her lexical diversity, while the second measure reflects the child's ability to use linguistic devices (cohesive ties) to maintain textual cohesion. Both of these measures were found to be good predictors of age and demonstrated significant age-related differences in To's study (in press). To ensure reliable and valid scoring, the target vocabulary and connectives (based on adult's production) can be given in the score sheet and examiners are only required to recognize the items. This recognition method is likely to ensure a high inter-rater reliability in scoring.

Employing Kindergarten Teachers as the CEIT's Administrator

The screening tool under investigation is the first locally published language screening test uniquely using kindergarten teachers and special child care workers (SCCWs) as test administrators. Teachers were viewed as good raters of children's language ability with high ecological validity and cost-effectiveness (Illerbrun, Haines and Greenough, 1985). However, due to time and resource constraints, the present study was not able to recruit a sufficient number of kindergarten teachers as blind examiners. Despite this, from the investigator's and her colleagues' experience during test administration and scoring, some suggestions can be made to enhance the effectiveness of using kindergarten teachers or SCCWs as test administrators. First, kindergarten teachers or SCCWs should be provided training and be tested to ensure fidelity in administration and scoring prior to their actual use of the CEIT. Training and testing can be delivered via a VCD/DVD along with the test for cost-effective reason. Only teachers who scored over a certain criteria, e.g. 80% accuracy in the test should be qualified to administer the CEIT to ensure reliable and valid screening results. This is necessary to ensure sensitivity and specificity of the CEIT---avoid over and under-referral.

Research Implications

Results from the present study suggest that CEIT was a reliable and valid screening tool with a high accuracy in identifying 5;01-5;06 children with LI. However, the study should be expanded to other normative age groups in CEIT, i.e. 3;00-6;00, with a use of a larger sample size. If possible, the reference standard used in the future study should

incorporate clinical impression rather than using a single test score to reduce effects from performance factors. Besides, an appropriate spectrum of children should be recruited to investigate the accuracy of the screening test since the children's diagnostic status in real-time clinical situations was less clear cut (Klee, 2008).

Further research on the CEIT should investigate the inter-rater reliability between speech therapists and kindergarten teachers or SCCWs. Besides, investigation of any curriculum effect of the screening test can be another research direction. The CEIT was decided to help all teachers in all local kindergartens and nurseries for language screening purpose. Therefore, feasibility and applicability of this screening test in identifying at-risk children in non-Po Leung Kuk kindergarten or nursery should be further evaluated.

References:

Anderson, L.M., Shinn, C., Fullilove, M.T., Scrimshaw, S.C., Fielding, J.E., Norman, J., et al. (2003). The effectiveness of early childhood development programs: A systematic review. *American Journal of Preventive Medicine, 24(3 supp.)*, 32–46.

Bossuyt, P. M., Reitsma, J. B., Bruns, D. E., Gatsonis, C. A., Glasziou, P. P., Irwig, L. M., et al. (2003). Towards a complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. *British Medical Journal, 326*, 41–44.

Cliff, N. (1993). Dominance statistics: Ordinal analyses to answer ordinal questions. *Psychological Bulletin, 114,* 495-509.

Dollaghan, C.A. (2007). *The handbook for evidence-based practice in communication disorders.* Baltimore: Paul H. Brookes.

Fletcher, P., Leonard, L. B., Stokes, S. F., & Wong, A. M-Y. (2005). The expression of aspect in Cantonese-speaking children with Specific Language Impairment. *Journal of Speech, Language and Hearing Research, 48*, 621-634.

Fletcher, P., Leung, C.-S. S., Stokes, S., & Weizman, Z. (2000). *Cantonese Pre-school Language Development: A Guide* (Report of the project "Milestones in the learning of spoken Cantonese by pre-school Children"). Hong Kong: Language Fund.

Fong, Y. Y. (2007). *Validation of a Screening Tool for Identifying Three-Year-Old Cantonese-Speaking Children with Language Impairment.* Hong Kong: University of

Hong Kong.

Hong Kong Society for Child Health and Development (987). *Manual of the Reynell Developmental Language Scales, Cantonese (Hong Kong) Version.* Hong Kong: Author.

Illerbrun, D., Haines, L., & Greenough, P. (1985). Language identification screening test for kindergarten: A comparison with four screening and three diagnostic language tests. *Language, Speech, and Hearing Services in Schools, 16*, 280-292.

Klee, T. (2008). Considerations for appraising diagnostic accuracy studies in communication disorders. *Evidence-Based Communication Assessment and Intervention, 2,* 34-45.

Klee, T., Wong, A. M-Y., Stokes, S. F., Fletcher, P., & Leonard, L. B. (2009). Assessing Cantonese-speaking children with language difficulties from the perspective of evidence-based practice: current practice and future directions. In S. P. Law, B. S. Weekes. & A. M-Y. Wong (Ed.), *Language Disorders in Speakers of Chinese* (pp. 89-111). Bristol, UK; Buffalo NY: Multilingual Matters.

Lee, P. W. H., Luk, E. S. L., Yu, K. K., & Bacon-Shone, J. H. (1985). A developmental language screening scale for use in Hong Kong. *The Hong Kong Journal of Paediatrics, 2*, 152-165.

Leonard, L. B., Wong, A. M-Y., Deevy, P., Stokes, S. F., & Fletcher, P. (2006). The production of passives by children with specific language impairment acquiring English or Cantonese. *Applied Psycholinguistics, 27,* 267-299.

Lin, Baogui. (1994). Yu Yan Zhang Ai Yu Jiao Zhi. Taipei: Wu-Nan Culture Enterprise.

Lund, N., & Duchan, J. (1993). *Assessing children's language in naturalistic contexts* (3rd ed.). Englewood Cliffs, NJ: Prentice-Hall.

McAlister, F. A., Straus, S. E., & Sackett, D. L. (1999). Why we need large, simple studies of the clinical examination: The problem and a proposed solution. *Lancet, 354*, 1721-1724.

McCauley, R. J., & Swisher, L. (1984). Psychometric review of language and articulation tests for preschool children. *Journal of Speech and Hearing Disorders, 49,* 34-42.

Paul, P. (2007). *Language disorders from infancy through adolescence*. (3rd ed.). St. Louis: Mosby.

Plante, E., & Vance, R. (1994). Selection of preschool language tests: A database approach. *Language, Speech, and Hearing Services in Schools, 25*, 15-24.

Po Leung Kuk District-based Speech Therapy Team. (2008). *Cantonese Oral Language Deficiency Early Identification Test for Pre-primary Children*. Hong Kong: Quality Education Fund.

Romano, J., Kromrey, J. D., Coraggio, J., & Skowronek, J. (2006). *Appropriate Statistics for Ordinal Level Data: Should We Really be Using t-test and Cohen's d for Evaluating Group Differences on the NSSE and Other Survey?* Paper presented at the annual meeting of the Florida Association of Institutional Research, February 1-3, 2006, Cooca Beach, Florida.

Scott, C. M. (2004). Syntactic ability in children and adolescents with language and learning disabilities. In R. A. Berman (Ed.), *Language Development across Childhood and Adolescence. Amsterdam.* Philadelphia: John Benjamins Pub.

Stokes, S. F., & So, L. K. H. (1997). Classifier use by language-disordered and age matched Cantonese-speaking children. *Asia Pacific Journal of Speech, Language and Hearing, 2(1)*, 83-101.

Tse, S. K. (2006). *The Oral Language Development of Hong Kong Preschool Children*. Hong Kong, Hong Kong University Press.

T'sou, B., Lee, T., Tung, P., Chan, A., Man, Y., & To, C. (2006). *Hong Kong Cantonese Oral Language Assessment Scale.* Hong Kong: City University of Hong Kong Press.

Whiting, P., Rutjes, A. W. S., Reitsma, J. B., Bossuyt, P. M. M., & Kleijnen, J. (2003). The development of QUADAS: A tool for the quality of assessment of studies of diagnostic accuracy included in systematic reviews. *BMC Medical Research Methodology, 3*, 25.

Wong, V., Lee, P.W.H., Lieh-Mak, F., Yeung, C.Y., Leung, P.W.L., Luk, S.L., et al. (1992). Language screening in preschool Chinese children. *European Journal of Disorders of Communication, 27*, 247-264.

To, C.K.S., Stokes, S.F., Cheung, H.T., & T'sou, B. (in press). Narrative assessment for Cantonese-speaking children. *Journal of Speech and Hearing Services in Schools*.

Appendix A

Case History Form

Background Information 基本資料

Child's Name 學童姓名: _____ Today's Date 填寫日期: _____

Sex 學童性別: _____ Date of Birth 出生日期: _____

Age 年齡: _____years 歲 _____months 個月

就讀幼稚園: _____(年級:_____) 課程: __一般學位/ 兼收學位/ EETC___

Respondent 填寫人姓名: _____

Relationship with Child 與學童之關係: _____ Tel 聯絡電話: _____

Please check (√) all appropriate items 請在適當的空格內填上 "√" 號

Child Background 學童背景

1. Has your child ever received a speech and language assessment?
   貴子弟曾否接受言語評估?
   No 沒有 □ Yes 有 □ (Time 時間:_____ Location 地方:_____
   Results 結果: _____)
2. Has your child ever received speech and language therapy?
   貴子弟曾否接受言語治療?
   No 沒有 □ Yes 有 □ (Time 時間:_____ Location 地方:_____
   Reason 原因: _____
   Completion of therapy 現在是否仍然接受治療? _____)
3. Has your child ever been diagnosed to have any other problems (e.g. sensory deficits, ear infection, autistic spectrum disorder, mental retardation or attention deficits)?
   貴子弟曾否被診斷患有其他病患 (例: 感覺障礙,中耳炎,自閉症,智障,專注力失調)?
   No 沒有 □ Yes 有 □ (Please specify 請註明: _____)
4. Do you have any concerns about your child's speech and language development?
   您對貴子弟的言語發展有沒有顧慮?
   No 沒有 □ Yes 有 □ (Please specify 請註明: _____)

Family Background 家庭背景

1. What language(s) are used at home?
   在家中使用哪種語言 (可多於一種)? _____
2. How much time will you use to speak to your child in Cantonese?
   您會花多少時間與貴子弟以廣東話溝通?

never 從不					half 一半					always 經常
0 □	1 □	2 □	3 □	4 □	5 □	6 □	7 □	8 □	9 □	10 □

Appendix B

QUADAS checklist for assessing the quality of study of diagnostic accuracy

Items		Yes	No	Unclear
1.	Was the spectrum of patients representative of the patients who will receive the test in practice?	( )	( )	( )
2	Were selection criteria clearly described?	( )	( )	( )
3	Is the reference standard likely to correctly classify the target condition?	( )	( )	( )
4	Is the time period between reference standard and index test short enough to be reasonably sure that the target condition did not change between the two tests?	( )	( )	( )
5	Did the whole sample or a random selection of the sample, receive verification using a reference standard of diagnosis?	( )	( )	( )
6	Did patients receive the same reference standard regardless of the index test result?	( )	( )	( )
7	Was the reference standard independent of the index test (i.e. the index test did not form part of the reference standard)?	( )	( )	( )
8	Was the execution of the index test described in sufficient detail to permit replication of the test?	( )	( )	( )
9	Was the execution of the reference standard described in sufficient detail to permit its replication?	( )	( )	( )
10	Were the index test results interpreted without knowledge of the results of the reference standard?	( )	( )	( )
11	Were the reference standard results interpreted without knowledge of the results of the index test?	( )	( )	( )
12	Were the same clinical data available when test results were interpreted as would be available when the test is used in practice?	( )	( )	( )
13	Were uninterpretable/ intermediate test results reported?	( )	( )	( )
14	Were withdrawals from the study explained?	( )	( )	( )

Source: Available in Whiting et al. (2003)

Appendix C

Psychometric Criteria

1	Definition of the standardization population, including geographic residence, socioeconomic status, and the "normalcy" of the sample.
2	Normative sample sizes of 100 or more per age or grade subgroup.
3	Evidence that item analysis, including analysis of item difficulty and item validity, were conducted during the development of the test.
4	Report of means and standard deviations (or their equivalent). To pass, these had to be provided for every age or grade subgroup.
5	Evidence of concurrent validity. This was taken as a significant correlation with other language tests of similar content.
6	Evidence of predictive validity. Acceptable evidence included a significant correlation or association between the test score and another measure of language given at a later time, or an association between test scores and need for treatment at a later time.
7	Estimate of test-retest reliability, including a correlation coefficient of .90 or better that was significant at the .05 level.
8	Estimate of inter-examiner reliability, including a correlation coefficient of .90 or better that was significant at the .05 level.
9	Test administration described in sufficient detail to ensure replication of the administration and scoring procedures on which the test norms are based.
10	Tester qualifications, including general background and specialized training, required for those administering and scoring the test.

Source: Available in McCauley and Swisher (1984); Plante and Vance (1994)

Acknowledgements

I would like to thank Dr. Anita Wong for the great support and advice throughout this dissertation project. Thanks also given to Miss Pensy Cheung for her generous sharing of experience in developing PLK preschool language test. Also, I gratefully acknowledge the help offered by Miss Carole Kwok, Miss Kathleen Lau, Miss Carmen Chong, Mr. Donald Lam and Mr. Man Ho in the data collection and analysis processes. Last but not least, thanks are given to the principals, teachers, parents and children of the following nurseries EETC for their help and cooperation during data collection:

Po Leung Kuk Ping Shek Kindergarten-cum-Nursery

Po Leung Kuk Fong Tam Yuen Leung Kindergarten-cum-Nursery

Hong Chi Pinehill Pre-school Centre (EETC)

Finally, my wholehearted gratitude is given to my Lord Jesus Christ for thy grace.