THE JOURNAL OF ASIA TEFL Vol. 5, No. 2, pp. 117-140, Summer 2008

Validation of the C-Test amongst Chinese ESL Learners[*]

Lei Lei
Huazhong University of Science and Technology, China

This paper reports on the validation of the C-Test amongst Chinese ESL learners. The hypothesis formulated is that the C-Test is a measure of general language proficiency. A total of 265 Chinese tertiary ESL learners participated in the study. Results from item discrimination analysis, internal consistency analysis, correlation analysis and confirmatory factor analysis confirmed that the C-Test has satisfactory item discrimination and is a reliable and valid measure of general language proficiency amongst Chinese ESL learners. I propose that two texts of the C-Test battery be chosen with its difficulty appropriate to the testees' proficiency and three beyond their proficiency in order to achieve satisfactory item discrimination. As for its pedagogical implications, we propose that the C-Test be utilized as a placement test and an anchor test in test equating.

Key words: C-Test, Validation, Chinese ESL learners

INTRODUCTION

The issue on what the C-Test actually measures has become a topic of lively debate in recent years. Many studies propose that the C-Test is a

[*] This article is sponsored by HUST Humanities and Social Sciences Research Project for Young Researchers (Project No. 2006001) and Humanities in Scientific and Technological Development: Project 985.

measure of general language proficiency with significant reliability and validity (e.g., Dörnyei & Katona, 1992; Eckes & Grotjahn, 2006; Klein-Braley, 1994, 1997; Raatz & Klein-Braley, 2002). However, other researchers challenge the C-Test for its lack of face validity, poor item discrimination and unclear construct validity (Bahaii & Ansary, 2001; Bradshaw, 1990; Cleary, 1988; Feldmann & Stemmer, 1987; Hood, 1990; Jafarpur, 1995; Kamimoto, 1992; Sigott & Kobrel, 1996). Despite the bulk of findings reports, the nature of the C-Test has not yet been fully investigated in the Asian context (e.g., Sabariah, 2002). Thus, the present study aims to contribute to this line of research and validate the C-Test amongst Chinese ESL learners.

LITERATURE REVIEW

Construction and Format of the C-Test

A C-Test usually consists of four to six authentic and thematically distinct texts and a total of at least 100 mutilated items. The texts are ordered in the C-Test in increasing difficulty levels. In terms of its construction, it employs the principle of "rule of two": it deletes the second half of every other word starting from the second word of the second sentence (Klein-Braley, 1985; Klein-Braley & Raatz, 1984; Raatz & Klein-Braley, 1981). If a word has an odd number of letters, the larger "half" is deleted. The first sentence of each text is usually left standing and proper names, numbers and one-letter word are undamaged. Through the texts, the mutilated part of each word was indicated by a single underline of constant length.

An example of English C-Test text is illustrated as follows:

> A popular form of recreation in Britain is attendance at dog racing. The fi___ impression o___ the ar___ is attra___. However, t___ races thems___ are uninte___ - a f___ dogs cha___ a tin ha___ - but thi___-two mil___ people att___ them annu___. Out o___ two ho___ barely fi___ to t___ minutes a___ usually dev___ to t___ actual rac___. There wo___ be

> n___ interest i___ it if it were not for the betting. Many of the audience pay little attention to the racing, but have their eyes fixed on a board which gives the number of the winners.(in Grotjahn, Klein-Braley and Raatz 2002, p. 96)

C-Test as a Measure of General Language Proficiency

A C-Test is an integrative written test of general language proficiency devised on the concept of reduced redundancy (Raatz & Klein-Braley, 2002). The reduced redundancy principle starts with the assumption that educated adult native speakers could draw on the redundancy of language to restore the damaged messages by their language competence (Oller, 1976; Spolsky et al., 1968). However, it is not the redundancy of text that is measured, but the examinee's ability to draw on the general language redundancy as a whole in order to restore the damaged text (Raatz & Klein-Braley, 2002). Consequently, the C-Test is assumed as an operationalization of the reduced redundancy principle and the way in which the examinees perform under these conditions is then used as an indicator of their general language proficiency (Eckes & Grotjahn, 2006; Klein-Braley, 1994, 1997). Put in other words, the C-Test, like the classic cloze test, is argued as a global assessment of general language proficiency (e.g., Dörnyei & Katona, 1992).

As Raatz and Klein-Braley (2002) have stated, the concept of general language proficiency tested by the C-Test seems to be similar, to a great extent, to what Bachman and Palmer (1982) have defined as operational competence or what Bachman (1990, p. 87) has termed as organizational competence, which is further divided into grammatical competence and textual competence. Thus, it seems to suggest that what the C-Test measures is not only lexical, morphological, syntactic, graphological knowledge on the sentence level, but also knowledge of cohesion and rhetorical organization on the text level.

However, the general language proficiency is by no means "a single, psychological simple construct" (Eckes & Grotjahn, 2006), i.e. the unidimensionality of general language proficiency does not imply that the testee's performance

on the items in a test is due to a single psychological process (Bejar, 1983, p. 31, cited in Eckes & Grotjahn, 2006). As Bachman (1990, p. 68) puts it, "it is now generally agreed that language proficiency is not a single unitary ability, but that it consists of several distinct but related constructs in addition to a general construct of language proficiency". As a consequence, what the C-Test requires of the testee's restoring the mutilated language is "the integration of both skills and knowledge: a core ability in all kinds of language use." (Eckes & Grotjahn 2006)

The findings of many studies lend supports to the C-Test as a measure of the general language proficiency. They demonstrate psychometrically significant correlation between the C-Test and other tests as well as such test components as vocabulary, grammar, reading, writing and even listening and speaking (Babaii & Ansary, 2001; Chapelle & Abraham, 1990; Daller & Phelan, 2006; Dörnyei & Katona, 1992; Eckes & Grotjahn, 2006; Farhady & Jamali, 1999; Hastings, 2002; Japarpur, 2002).

The C-Test can be used for first, second and foreign language testing (Grotjahn, Klein-Braley, & Raatz, 2002) and have been utilized in various situations and for a variety of purposes.

Merits and Criticisms of the C-Test

Developed by Raatz and Klein-Braley in 1981, the C-Test was proposed as an alternative to the classic cloze test, primarily in response to the criticism of numerous defects upon the cloze test: 1) the cloze test is relatively long, 2) the cloze test usually consists of only one longer text, which results in test bias in terms of text specificity, 3) the factors of "text", "deletion rate" and "starting point" affect reliability and validity coefficients and the difficulty of the cloze test, 4) the paradox of exact or acceptable scoring exists and 5) many cloze tests reported in the literature are less reliable than originally assumed (Grotjahn, Klein-Braley, & Raatz, 2002; Raatz & Klein-Braley, 2002). Nevertheless, the C-Test resembles the overall format and appearance of the classic cloze test. As Klein-Braley (1997, p. 63) puts it, "C in the name

C-Test was chosen specifically as an abbreviation of the word 'cloze' in order to indicate the relationship between the two test procedures. The C-Test was an attempt to retain the positive aspects of cloze tests but to remedy their technical defects."

However, the C-Test has acclaimed to be superior to the classic cloze test in its economy in construction (following the "rule of two", it is trouble-free to construct a C-Test), administration (less than 30 minutes to administer a C-Test) and scoring (around 1 or 2 minutes to score a C-Test) as well as its high reliability and validity (Dörnyei & Katona, 1992; Eckes & Grotjahn, 2006; Klein-Braley, 1997; Klein-Braley & Raatz, 1984). Thanks to its salient advantages, the C-Test has been utilized in various situations and for a variety of purposes, e.g., as a placement test, as an anchor test in test equating, and as a research instrument in cognitive and applied linguistics, etc. (Eckes & Grotjahn, 2006; Grotjahn, Klein-Braley, & Raatz, 2002).

Despite its merits in various aspects, the C-Test has not been without its critics. Bahaii and Ansary (2001) summarize the flaws concerning the C-Tests as: lack of face validity, poor item discrimination and unclear construct validity (Bradshaw, 1990; Cleary, 1988; Feldmann & Stemmer, 1987; Hood, 1990; Jafarpur, 1995; Kamimoto, 1992; Sigott & Kobrel, 1996).

First of all, the C-Test is often leveled at its low face validity. Teachers or students incline to interpret the C-Test as a test of reading comprehension or a special form of intelligence tests. They hesitate to accept it as a test of integrative language proficiency. It is the very reason that Raatz and Klein-Braley (2002) propose it be necessary to invest some time in public relations work in advance of its use.

Moreover, the C-Test has also been censured for its being too easy as a language proficiency test (Cleary, 1988; Jafarpur, 2002), which may lead to poor item difficulty or item discrimination and ceiling effect may occur in such case. Another problem which needs caution is that when it is too easy, it involves too much automaticity and lack of conscious control of the examinees (Babaii & Moghaddam, 2006; Green, 1998). In such case, the validity of the C-Test as an integrative measure of general language

proficiency is questioned while the examinees restore the mutilated items merely by drawing on their knowledge of micro-level cues, e.g., grammatical and/or morphological features rather than by understanding the whole message adequately (Bahaii & Ansary, 2001). In response to the problem of low discrimination, Cleary (1988) proposes that the C-Test item discrimination could be enhanced by left-hand deletions. Furthermore, Kamimoto (1993) suggests tailoring the C-Test through classical item analysis in order to improve the discrimination (cf. Jafarpur (2002) found that the classical item analysis approach showed no gains in the discrimination values). In addition, Sigott and Kobrel (1996) recommend increasing the text difficulty by a bigger proportion of deletion (deleting two third of the words instead of a half) and left-hand deletion in Cleary's (1998) terms.

Finally, the most controversial issue regarding the C-Test is what it actually measures. Some dispute that the C-Test is more of a measure of grammatical and reading ability and primarily of micro-level skills, and less of a measure of textual competence and macro-level skills (Chappelle & Abraham, 1990; Cohen et al., 1985; Kamimoto, 1992; Stemmer, 1991). However, many professionals argue that it is a measure of global proficiency in language (Dörnyei & Katona 1992; Hastings, 2002; Klein-Braley, 1994, 1997). In response to the argument that C-Test is primarily a measure of micro-level skills, Babaii and Moghaddam (2006) prove that employing texts with more syntactic complexity and abstraction will result in more difficult test tasks, which in turn encourages more frequent use of the testee's macro-level processing. Moreover, the stance of general language proficiency does not construe the language ability as a single or indivisible entity, rather, as Carroll (1993, p. 191) interprets, as a "measure of the extent to which an individual has learned the lexical and grammatical phenomena of a language" (cited in Eckes & Grotjahn, 2006).

Despite the bulk of findings discussed earlier, the studies do not provide much attention to exploration of the nature of the C-Test in the Asian context. To our knowledge, Sabariah (2002) is to date the only study on the validation of the C-Test amongst Asian ESL learners. Her study on a group of

secondary Malay male students suggests that the C-Test is a reliable measure of language proficiency and behaves like parallel forms of a test. Thus, the primary consideration of the study is to contribute to this line of research and validate of the C-Test amongst Chinese ESL learners.

RESEARCH HYPOTHESIS AND RESEARCH DESIGN

Research Purpose and Hypothesis

The aim of the study is to validate the C-Test amongst Chinese ESL learners. The hypothesis formulated based on the discussion above is that the C-Test is a measure of general language proficiency which consists of several distinct but related constructs.

Approaches of internal consistency analysis, correlation analysis, confirmatory factor analysis are employed in this study to verify the hypothesis. As for the investigation of the construct validity of the C-Test, a structural model is proposed, in which the C-Test will join all of the six sections (i.e., listening, reading, vocabulary, classic cloze, translation and writing) of an English test to constitute the manifest variables, with the aim to measure the latent variable (the general language proficiency). The structural model is illustrated in Figure 1:

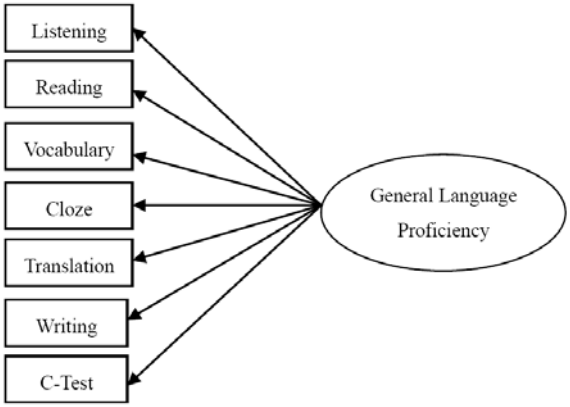


FIGURE 1
The Structural Model

Participants

A total of 265 freshmen non-English majors at Huazhong University of Science and Technology, China P.R., participated in the present research. Of them, 131 were male and 134 were female. All of the participants had passed the college matriculation examinations and had learned English as their second language for at least six years.

At the time the study was conducted, all of the participants were attending the compulsory *College English* programme eight hours a week. They were students of six natural classes of the *College English* programme randomly chosen by the researchers. They were required to finish Book I to Book IV of *College English (New Edition)* published by Shanghai Foreign Language Education Press in 2000 within two terms, two books each. The C-Test was administered near the end of their first term when they were finishing *College English* Book II.

Three weeks before the administration of the C-Test, the participants had taken the test for *College English* Book I (Band I). Four weeks after the administration, they sat another test for *College English* Book II (Band II).

Instruments

Two instruments were employed in the research, i.e. the C-Test battery and the score of the test for *College English* Book I (the Band I test hereafter).

The C-Test battery was developed by the researchers under the guidance of C-Test construction principles proposed by Klein-Braley and Raatz (1984) and Klein-Braley (1985). It consists of five texts with 20 blanks each and the length of the texts goes from 94 to 117 words. All texts were chosen from *21st Century College English* co-published by Higher Education Press and Fudan University Press in 1999 for the reason that the compilers claimed that all materials in the series of the course book were authentic texts written by native English speakers and appropriately elaborated for ESL college learners in China. The first three texts were paragraphs selected from Book I to Book III respectively and both the fourth and the fifth texts from Book IV of *21st Century College English.*

We bore the following principles in mind as well during the construction of the C-Test battery: 1) the texts were of a variety of themes and styles (except literary works); 2) the texts were ordered in an increasing degree of difficulty; 3) Text 3 to 5 were intentionally selected from Book III and Book IV and might be beyond some of the participants' English proficiency, with the aim of guaranteeing the item discrimination of the test and of avoiding the ceiling effect.

The participants were required to write down information concerned on the test sheet of the C-Test: their gender, student number and score on the College Matriculation English Test (the MET hereafter). The reason for so doing was that the student number would help locate their score on the Band I test and the MET was taken as the criterion measure testing the item discrimination of the C-Test and the concurrent validity of the Band I test. Of the 265 participants, 259 provided their score on the MET.

The Band I test was a test composed of six sections: listening comprehension (10 short dialogues with one item each and three short passages with 5 items each, altogether 20% of the total score), reading comprehension (four

passages with five items each, 40% of the total score), vocabulary (20 items, 10% of the total score), the classic cloze test (20 items, 10% of the total score), translation (translation of phrases in five Chinese sentences into English, 5% of the total score) and writing (writing of an argumentative essay of about 120 English words, 15% of the total score). The items of the first four sections were all objective multiple choice ones and those of the last two sections were subjective ones. The total score of the test was 100 points.

Administration

The administration procedure of the Band I test went on as follows: the test was administered in a standardized testing environment and participants were allowed 120 minutes to complete the test. They were first asked to listen to the tape recording with earphones and complete the listening comprehension section, which lasted about 20 minutes. Afterwards, the participants finished the sections of reading comprehension, vocabulary, classic cloze, translation and writing within 100 minutes.

In terms of scoring, the first four multiple choice sections were scored by machine and the translation and writing sections were scored twice by both the *College English* teachers of the participants and the researchers respectively. If different scoring results occurred, the researchers would discuss with the teachers and tried to reach an agreement.

The C-Test battery was administered within a classroom environment with the aid of the *College English* teachers. The whole process of completing the C-Test lasted about 30 minutes on average. To make the scoring procedure of C-Test standardized, spelling errors were counted as incorrect. Each correct response was awarded one point, thus 20 points for each text with a total of 100 points. The C-Test was scored by the researchers.

Data Analysis

The procedure of data analysis of this study involves the following six

steps:

First of all, descriptive statistics of the C-Test, the Band I test and the MET were computed to demonstrate the primary feature of them.

Next, biserial correlation coefficients, i.e. Pearson correlation coefficients between the texts in the C-Test battery and the MET, were computed to test the item discrimination of the C-Test. The theoretical foundation for so doing was that the biserial or Pearson correlation coefficients were to determine whether the attribute measured by the criterion was also measured by the test item and the extent to which the item measured them (Henning 2001, p. 51-56). The MET was taken as the criterion measure. Since the MET is a large-scale high-stake test with authority in China, it is reasonable to accept its discrimination. If the texts were satisfactorily correlated with the MET, we were safe to suppose that the C-Test had acceptable discrimination.

Moreover, reliability or internal consistency of the C-Test battery developed and used at the study was analyzed. One point which needs caution in the C-Test consistency analysis is that each C-Test text is taken as a superitem or testlet with item values corresponding to the number of blanks filled in correctly (Eckes & Grotjahn, 2006; Raatz & Klein-Braley, 2002). The reason for so doing is that the individual blanks in the texts are dependent on each other as a result of text structure and content, whereas the consistency analysis assumes that all items entering into the equation are independent of each other (Eckes & Grotjahn, 2006; Raatz & Klein-Braley, 2002). As for the present study, there were five superitems with a score between 0 and 20 since the C-Test battery consisted of five texts with 20 items each.

In addition, the concurrent validity of the Band I test was analyzed, i.e. the correlation coefficient (Pearson correlation) between the score of the MET and the total score of the Band I test was computed, with the MET as the criterion test. Since the Band I test was utilized in testing the concurrent and construct validity of the C-Test, a satisfactory validity of the Band I test must be achieved to guarantee its statistical acceptability.

Furthermore, the concurrent validity of the C-Test battery was investigated. The correlation coefficient between the total score of the C-Test, the scores of

the six sections and the total score of the Band I test was analyzed to testify the relation between the C-Test and the Band I test, i.e. the concurrent validity of the C-Test battery. For the above statistical analyses, the software SPSS was employed.

Finally, the construct validity of the C-Test was examined. Confirmatory factor analysis was adopted to verify the structural model proposed earlier in Fig. 1, which was the central concern of the study. As for the acceptability of parameter statistics in confirmatory factor analysis, it is suggested that the values of NFI, RFI, IFI, TLI, CFI be more than .90 and the value of RMSEA be less than .60 to achieve a statistically satisfactory goodness-of-fit (Byrne 2001). The software AMOS was utilized for this part of statistics.

RESULTS

Descriptive Statistics of the Instruments

Descriptive statistics of the C-Test, the Band I test and the MET are presented in Table 1. The results showed that the C-Test total score ranged from 25 to 80 with a mean 51.74, the Band I test total score from 41 to 92 with a mean 73.65 and the MET total score from 66 to 144 with a mean 118.91. As for the standard deviation (S.D.), those of the C-Test texts were similar in value, while those of the Band I reading and listening sections were much more than those of the other sections in value. The dissimilarity of reading and listening S.D.s to the others signified that the subjects of the study deviated in reading and listening much more than in other language skills or knowledge.

TABLE 1
Descriptive Statistics of the Instruments

	N	Min.	Max.	Mean	S.D.
Text1	265	4	16	10.21	2.21
Text2	265	4	19	11.05	2.92
Text3	265	1	17	8.97	2.93
Text4	265	1	19	12.09	3.21
Text5	265	0	18	9.43	3.17
C-Test total	265	25	80	51.74	10.26
Listening	265	4	20	14.55	3.23
Reading	265	10	38	28.56	5.21
Vocabulary	265	3	10	7.84	1.24
Classic cloze	265	1	9	6.31	1.42
Translation	265	1	15	4.25	1.03
Writing	265	8	15	12.15	1.29
Band1 total	265	41	92	73.65	9.54
MET total	259	66	144	118.91	11.51

Item Discrimination of the C-Test

The item discrimination statistics of the C-Test are reported in Table 2. The results showed that all texts in the C-Test battery were significantly correlated with the MET total at the .000 level. The findings implied that the C-Test had satisfactory discrimination.

TABLE 2
Item Discrimination Statistics of the C-Test

	Text 1	Text 2	Text 3	Text 4	Text 5
MET total	.325**	.290**	.198**	.239**	.247**

** significant at $p < .000$

Reliability of the C-Test

To assess whether the five texts that constituted the C-Test battery formed a reliable scale, Cronbach's alpha or internal consistency was computed. As

the results shown in Table 3, the alpha for the five texts or superitems was .747, which indicated that the C-Test had reasonable internal consistency reliability.

TABLE 3
Reliability Statistics of the C-Test

N	N of Items	Mean	Std. Deviation	Cronbach's Alpha
265	5	51.74	10.258	.747

Concurrent Validity of the Band I Test

To testify the concurrent validity of the Band I test, we took the MET as the criterion measure of the Band I test and computed the Pearson correlation between the MET and the Band I test. The result showed that the correlational coefficient was .476 ($p < .000$), which signified that the score of the Band I test was a valid indicator of the participants' English language proficiency.

Concurrent Validity of the C-Test

The correlation coefficients between the total score of the C-Test, the scores of the six sections and the total score of the Band I test were computed in order to testify the concurrent validity of the C-Test. The results are referred to in Table 4.

TABLE 4
Concurrent Validity of the C-Test

	Listening	Reading	Vocabulary	Cloze	Translation	Writing	Band I total
C-Test	.318*	.317*	.338*	.408*	.258*	.348*	.461*

* significant at $p < .000$

The results showed that the C-Test was significantly correlated with the scores of all sections and the total score of the Band I test ($p < .000$). The

highest correlation was with the total score of the Band I test (.461), followed by the classic cloze test (.408), writing (.348) and vocabulary (.338) sections. The least correlated was with the translation section (.258).

Construct Validity of the C-Test

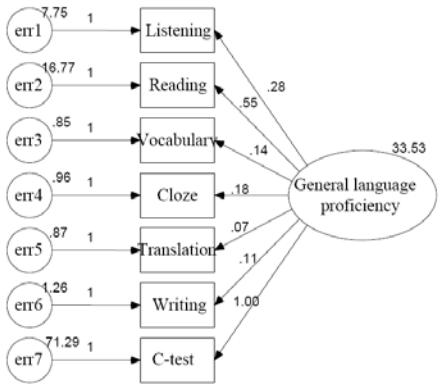
The statistics of confirmatory factor analysis in testing the construct validity of the C-Test are presented in Table 5.

TABLE 5
Statistics of Confirmatory Factor Analysis

X^2	X^2/df	NFI	RFI	IFI	TLI	CFI	RMSEA
23.150	1.654	.940	.910	.975	.962	.975	.050

It is safe to conclude, based on the statistics in Table 5, that the data within the model fit well and the manifest variables defined the latent variable (general language proficiency) well. Therefore, the model was statistically significant and accordingly accepted. The output path diagram is illustrated in Figure 2.

FIGURE 2
The Output Path Diagram



The standardized factor loadings (i.e. the standardized regression weights) of the Band I test sections and the C-Test are shown in Table 6. Results showed that all the loadings were significant at $p < .001$ level. The loading of classic cloze factor was the highest observed (.721), which indicated that the classic cloze factor contributed the most highly to the explanation of the latent variable (general language proficiency). The C-Test loading (.566) ranked in weight after vocabulary (.667) and reading (.616), followed by listening (.506), writing (.485) and translation (.422).

TABLE 6
Standardized Factor Loadings of the Band I test Sections and the C-Test

Listening	Reading	Vocabulary	Cloze	Translation	Writing	C-Test
.506**	.616**	.667**	.721**	.422**	.485**	.566**

** significant at $p < .001$

DISCUSSION

There are several aspects of the results that have to be addressed. The first problem deals with the item discrimination of the C-Test. It seems to be reasonable to assume that the C-Test has significant item discrimination amongst the Chinese EFL learners at the study. The problem of low discrimination often arises when the C-Test is either too easy (floor effect) or too difficult (ceiling effect). Either Cleary's (1988) proposal of left-hand deletion in the C-Test construction, or Kamimoto's (1993) suggestion of tailoring the C-Test through classical item analysis, or Cleary's (1998) methods of increasing the C-Test difficulty in case of its being too easy attempts to ensure the appropriate difficulty of the C-Test for the testees. In the construction the C-Test at the study, we used two texts which suited the participants' English proficiency and the other three which might be beyond some of the participants' proficiency. The choice of the texts with a degree of difficulty at the study seems to have explained why the C-Test has significant item discrimination. Accordingly, text difficulty corresponding to the testees

is important in the construction of the C-Test. In order to achieve satisfactory item discrimination, we propose that we have acquaintance of the testees' proficiency before the construction of the C-Test and choose two texts with its difficulty appropriate to the testees' proficiency and three beyond their proficiency.

The next problem relates to the internal consistency of the C-Test. The internal consistency coefficient suggests that the C-Test is a reliable measure. The findings provide comparable results with both Dörnyei and Katona (1992), which validated the C-Test on groups of university and secondary school Hungarian EFL learners with satisfactory reliability coefficients (.75 and .77), and Sabariah (2002), on a group of secondary Malay male students, which yielded reliability coefficients ranging from .76 to .86. The satisfactory internal consistency lends support to the fact that the C-Test is a reliable measure of the general English language proficiency amongst Chinese ESL learners.

The third aspect involves the concurrent validity of the C-Test. The results at the study showed that the C-Test was significantly correlated with the Band I test and all its sections. The highest correlation was with the Band I test, followed by the classic cloze test, writing, vocabulary and translation in degree of correlation. As discussed earlier, many studies have found significant correlation between the C-Test and other tests as well as testing components such as vocabulary, grammar, reading, writing, listening and even speaking (Babaii & Ansary, 2001; Chapelle & Abraham, 1990; Daller & Phelan, 2006; Dörnyei & Katona, 1992; Farhady & Jamali, 1999; Hastings, 2002; Japarpur, 2002). Sabariah (2002) in her study also found a satisfactory correlation between C-Tests and classic cloze tests and suggested that the C-Test appears to measure the same thing as what the classic cloze test is measuring, that is general English language proficiency. The results obtained at the study seemed to have supported the previous findings.

What of most interest to us was that the C-Test was correlated much more highly with the Band I test and the classic cloze test than with all the other sections. It appears to indicate that what the C-Test measures is the same as

what the test and the classic cloze are measuring, i.e. general English language proficiency. As Sabariah (2002) argues, the significant correlation between the C-Test and the cloze test signifies that the "C-Test provides as good estimates of language proficiency as the cloze, if not better" (Dörnyei & Katona, 1992, p. 193).

The last aspect applies to the construct validity of the C-Test. The proposed model that the C-Test along with all sections in the Band I test would significantly define general language proficiency is confirmed at the study. Similar results have been yielded from a confirmatory factor analysis by Eckes and Grotjahn (2006), in which the C-Test together with reading, listening, writing and speaking in a TestDaF defines general language proficiency. What is divergent in Eckes and Grotjahn (2006) is that the C-Test loadings are by far the highest observed (always exceeding .80). Nevertheless, the hypothesized model at the present study is significantly acceptable, which confirms the findings of Eckes and Grotjahn (2006) that the C-Test is a measure of general language proficiency.

As we have discussed earlier, the language proficiency is never a single or psychologically simple construct. Rather, it is divisible and is "a complex theoretical construct composed of various, more specific constructs" (Eckes & Grotjahn, 2006). The fact that the general language proficiency at the study is defined by a composite of C-Test and other abilities and skills bolsters the divisible point of view. From the discussion, one may conclude that the C-Test measures general language proficiency, which is an integration of various distinct but related constructs.

To summarize the discussion, the C-Test is a reliable and valid measure of the general language proficiency with acceptable item discrimination amongst the Chinese ESL learners. Thanks to its economy in construction, administration and scoring as well as its satisfactory discrimination, reliability and validity discussed above, the pedagogical implications lie in that the C-Test be utilized as a placement test in the classroom environment, an anchor test in test equating and a research instrument in cognitive and applied linguistics, etc.

CONCLUSION

The study intends to validate the C-Test amongst Chinese ESL learners. The hypothesis formulated is that the C-Test is a measure of general language proficiency which consists of several distinct but related constructs. Therefore, the main issue of concern is whether the C-Test is a reliable and valid measure of language proficiency amongst Chinese ESL learners.

The following conclusion seems to be sufficiently substantiated based on the findings of the study: results of item discrimination analysis and internal consistency analysis suggest that the C-Test has satisfactory item discrimination and is a reliable measure. Furthermore, the significant correlation between the C-Test and the test as well as all its sections used for the study seems to prove that the C-Test holds acceptable concurrent validity and measures the same thing as what the test and the classic cloze test are measuring, i.e. general language proficiency. Finally, the results of construct validity analysis lend further support to the claim that the C-Test is a valid measure of general language proficiency amongst Chinese ESL learners. One point concerning the construction of the C-Test is proposed as well. In order to achieve satisfactory item discrimination of the C-Test, we propose that we have acquaintance of the testees' proficiency before the construction of the C-Test and choose two texts with its difficulty appropriate to the testees' proficiency and three beyond their proficiency. As for its pedagogical implications, we propose that the C-Test be utilized as a placement test in the classroom environment, an anchor test in test equating and a research instrument in cognitive and applied linguistics.

ACKNOWLEDGEMENTS

I would like to thank Prof. Ulrich Raatz, University of Duisburg, Germany, for his generous help. Special thanks are also due to Prof. Yan Jin, Shanghai Jiao Tong University, China, for her invaluable suggestions in the preparation

of the paper.

THE AUTHOR

Lei Lei is lecturer at School of Foreign Languages, Huazhong University of Science and Technology, China, P.R. and is now doing his PhD research in SLA at Shanghai Jiao Tong University, China, P.R.. His current research interests cover SLA, psycholinguistics, and language testing. His recent publications include *A Study of Incidental Vocabulary Acquisition through Writing by Chinese EFL Learners* with Y. Wei, L. Ye and M. Zhang (2007) and *An Empirical Study on Writer's Block of Chinese EFL Learners at the Tertiary Level* with Y. Wei (2007).
Email: leileicn@hust.edu.cn

REFERENCES

Babaii, E., & Ansary, H. (2001). The C-Test: A valid operationalization of reduced redundancy of principle. *System, 29*(2), 209-219.

Babaii, E., & Moghaddam, M. J. (2006). On the interplay between test task difficulty and macro-level processing in the C-Test. *System, 34*(4), 586-600.

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford: Oxford University Press.

Bachman, L. F., & Palmer, A. S. (1982). The construct validation of some components of communicative proficiency. *TESOL Quarterly, 16*(4), 449-665.

Bradshaw, J. (1990). Test takers' reaction to a placement test. *Language Testing, 7*(1), 13-30.

Byrne, B. M. (2001). *Structural equation modeling with AMOS: Basic concepts, applications, and programming.* Mahwah, NJ: Lawrence Erlbaum Associates.

Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor-analytic studies.* Cambridge: Cambridge University Press.

Chappelle, C., & Abraham, R. (1990). Cloze method: What difference does it make? *Language Testing, 7*(2), 121-146.

Cleary, C. (1988). The C-Test in English: Left-hand deletions. *RELC Journal, 19*(1),

26-38.

Cohen, A. D., Segal, M., & Bar-Siman-Tov, R. W. (1985). The C-Test in Hebrew. In C. Klein-Braley, & U. Raatz (Eds.), *C-Tests in der praxis [C-Testing in practice]*. (pp. 121-127). Bochum: AKS-Verlag.

Daller, H., & Phelan, D. (2006). The C-test and TOEIC as measures of students' progress in intensive short courses in EFL. In R. Grotjahn. (Ed.). *Der C-Test: Theorie, empirie, anwendungen/ The C-test: Theory, empirical research, applications*. (pp. 101-109). Frankfurt am Main: Peter Lang.

Dörnyei, Z., & Katona, L. (1992). Validation of the C-test amongst Hungarian EFL learners. *Language Testing, 9*(2), 187-206.

Eckes, T., & Grotjahn, R. (2006). A closer look at the construct validity of C-Tests. *Language Testing, 23*(3), 290-325.

Farhady, H., & Jamali, F. (1999). Varieties of C-test as measures of general proficiency. *Journal of the Faculty of Foreign Languages, 3,* 23–42 (Tehran University Press).

Feldmann, U., & Stemmer, B. (1987). Thin___ aloud a___ retrospective da___ in c-te___ taking: Diffe___ languages — diff___ learners — sa___ approaches? In C. Faerch, & C. Kasper (Eds.). *Introspection in Second Language Research*. (pp. 251-267). Clevedon: Multilingual Matters.

Green, A. (1998). *Verbal protocol analysis in language testing research*. Cambridge: Cambridge University Press.

Grotjahn, R., Klein-Braley, C., & Raatz, U. (2002). C-Tests: An overview. In J. A. Coleman, R. Grotjahn, & U. Raatz (Eds.). *University Language Testing and the C-Test.* (pp. 93-114). Bochum: AKS-Verlag.

Hastings, A. J. (2002). In defense of C-Testing. In R. Grotjahn (Ed.). *Der C-Test: Theoretische Grundlagen und Praktische Anwundungen [The C-Test: Theoretical foundations and practical applications]*. Vol. 4. (pp. 11-25). Bochum: AKS-Verlag.

Henning, G. (2001). *A guide to language testing: Development, evaluation and research*. Beijing: Foreign Language Teaching and Research Press.

Hood, M. (1990). The C-Test: A viable alternative to the use of the cloze procedure in testing? In L. Arena (Ed.). *Language Proficiency*. (pp. 173-189). New York: Plenum Press.

Jafarpur, A. (1995). Is C-Testing superior to cloze? *Language Testing, 12*(2), 194-216.

Jafarpur, A. (2002). A comparative study of a C-Test and a cloze test. In R. Grotjahn (Ed.). *Der C-Test: Theoretische, grundlagen und praktische anwendungen [The C-test: Theoretical foundations and practical applications]*. Vol. 4. (pp. 31–51). Bochum: AKS-Verlag.

Kamimoto, T. (1992). An inquiry into what a C-Test measures. *Fukuoka Women's Junior College Studies, 44*, 67-79.

Kamimoto, T. (1993). Tailoring the test to fit the students: Improvement of the C-Test through classical item analysis. *Fukuoka Women's Junior College Studies, 30,* 47-61.

Klein-Braley, C. (1985). A cloze-up on the C-Test: A study in the construct validation of authentic tests. *Language Testing, 2*(1), 76-104.

Klein-Braley, C. (1994). *Language testing with the C-Test: A linguistic and statistical investigation into the strategies used by C-Test takers, and the prediction of C-Test difficulty*. Unpublished Higher Thesis, Department of Linguistics and Literature, University of Duisburg, Germany.

Klein-Braley, C. (1997). C-Tests in the context of reduced redundancy testing: An appraisal. *Language Testing, 14*(1), 47-84.

Klein-Braley, C., & Raatz, U. (1984). A survey of research on the C-Test. *Language Testing, 1*(2), 134-146.

Oller, J. W. Jr. (1976). Evidence for a general language proficiency factor: An expectancy grammar. *Die Neueren Sprachen, 75*, 165-74.

Raatz, U., & Klein-Braley, C. (1981). The C-Test – a modification of the cloze procedure. In T. Culhane, C. Klein-Braley, & D. K. Stevenson (Eds.). *practice and problems in language testing.* (pp. 113-48). Essex: University of Essex Occasional Papers.

Raatz, U., & Klein-Braley, C. (2002). Introduction to language testing and to C-Tests. In J. A. Coleman, R. Grotjahn & U. Raatz (Eds.). *University language testing and the C-Test.* (pp. 75-91). Bochum: AKS-Verlag.

Sabariah, M. R. (2002). Validating the C-test amongst Malay ESL Learners. In T.M.T. Mohtar, F. Haron, & S. Nackeeran (Eds). *Proceedings of Selected Papers of Fifth Malaysian English Language Teaching Association (MELTA) Biennial International Conference*, Malaysia: Petaling Jaya.

Sigott, G., & Kobrel, J. (1996). Deletion patterns and C-Test difficulty across languages. In R. Grotjahn (Ed.). *Der C-Test: Theoretische grundlagen und praktische anwundungen[The C-Test: Theoretical foundations and practical applications]*. Vol. 3. (pp. 159-172). Bochum: Brockmeyer.

Spolsky, B., Bengt, S. M., Sako, E. W. & Aterburn, C. (1968). Preliminary studies in the development of techniques for testing overall second language proficiency. In J. A. Upshur & J. Fata (Eds.). *Problems in foreign language testing, Language Learning Special Issue, 3*, 79-103.

Stemmer, B. (1991). *What's on a C-Test taker's mind: Mental process in C-Test taking*. Bochum: Brockmeyer.

APPENDIX

The C-Test Battery

Text 1

The revolution in communications is just beginning. It wi_____ take pl_____ over sev_____ decades, a_____ will b_____ driven b_____ new "applic_____" -- new to_____, often mee_____ currently unfore_____ needs. Dur_____ the ne_____ few ye_____, major deci_____ will ha_____ to b_____ made. I_____ is cru_____ that a br_____ set o_____ people -- not just technologists or those who happen to be in the computer industry -- participate in the debate about how this technology should be shaped. If that can be done, the highway will serve the purposes users want. Then it will gain broad acceptance and become a reality.

Text 2

People are generally prone to what language expert S. I. Hayakawa calls "the two-valued orientation." We ta_____ about see_____ both si_____ of a ques_____ as i_____ every ques_____ had on_____ two si_____. We ass_____ that ever_____ is eit_____ a suc_____ or a fail_____ when, i_____ fact, infi_____ degrees o_____ both a_____ possible. A_____ Hayakawa poi_____ out, the_____ a world of difference between "I have failed three times" and "I am a failure." Indeed, the words failure and success cannot be reasonably applied to a complex, living, changing human being. They can only describe the situation at a particular time and place.

Text 3

Exploring nature with your child is largely a matter of being open to what lies all around you. It i_____ learning ag_____ to u_____ your ey_____, ears, no_____ and finge_____, opening u_____ the dis_____ channels o _____ your sen_____. For mo_____ of u_____, knowledge o_____ our wo_____ comes lar_____ through si_____, yet w_____ look ab_____

with su_____ unseeing ey_____ that we are partially blind. One way to open your eyes to unnoticed beauty is to ask yourself, “What if I had never seen this before? What if I knew I would never see it again?”

Text 4

“I think the answer lies in that direction,” affirms Dr. Bridger. “Take the situation where someone is in a crisis. The Chi_____ word f_____ crisis I _____ divided in_____ two char_____, one mea_____ danger a_____ the ot_____ meaning oppor_____. We i_____ the Wes_____ world fo_____ only up_____ the ‘dan_____’ aspect o_____ crisis. Cri_____ in Wes_____ civilization h_____ come t_____ mean dan_____, period. And yet the word can also mean opportunity. Let us now suggest to the person in crisis that he cease concentrating so upon the dangers involved and the difficulties, and concentrate instead upon the opportunity—for there is always opportunity in crisis. Looking at a crisis from an opportunity point of view is a lateral thought.”

Text 5

By the time the children reach high school, something remarkable has happened. A sur_____ of t_____ children's par_____ and tea_____ found th_____ those w_____ as four-year-olds h_____ enough self-c_____ to ho_____ out f_____ the sec_____ marshmallow gene_____ grew u_____ to b_____ better adju_____, more pop_____, adventurous, conf_____ and depen_____ teenagers. T_____ children w_____ gave in to temptation early on were more likely to be lonely, easily frustrated and stubborn. They could not endure stress and shied away from challenges. And when some of the students in the two groups took the Scholastic Aptitude Test, the kids who had held out longer scored an average of 210 points higher.