http://e-flt.nus.edu.sg/

Electronic Journal of Foreign Language Teaching
2016, Vol. 13, No. 2, pp. 184–202

C-Tests in Vietnam: An Exploratory Study of English Proficiency

Elizabeth A. Hiser
(hiserelizabeth@gmail.com)
Massy University, New Zealand

Kiet Si Thang Ho
(kiet.ho@ufl.udn.vn)
University of Foreign Language Studies, University of Danang, Vietnam

Abstract

The purpose of this study is to investigate whether C-Tests, which have been tested with international ESL cohorts in New Zealand, can be used as reliable English proficiency measures in Vietnam. C-Tests produce robust reliability and validity in most SLA studies. The three C-Tests that have been used at Massey University, New Zealand, were used as a basis of comparison for the Vietnamese sample, as they have been trialled and revised using classical item analysis, reliability studies and construct/concurrent/criterion validity checks against IELTS/TOEIC scores in New Zealand, thereby considered reliable and valid for various Asian and Middle Eastern ethnic groups. The findings of this study show that the three C-Tests have acceptable reliability and significant correlations among themselves, and can be used to evaluate overall English proficiency in Vietnam. The results administered to the Vietnamese cohort in this study are expected to be added to the growing number of other ethnic groups for which they evaluate overall English proficiency validly, reliably, and efficiently. There are implications discussed here for test developers working with C-Tests in making them reliable and valid measures of English proficiency.

1 Introduction

C-Tests have been shown to produce robust reliability and validity in most SLA studies (Alderson, 2002; Daller & Phelan, 2006; Eckes & Grojahn, 2006; Hiser, 2012; Ikeguchi, 1998b). They are also tests that have simple, fast, and efficient scoring with selected items; they are easily created and quickly analysed for reliability and validity (Raatz & Klein-Braley, 2002). Advantages of C-Tests are that they can be used as a reference to give an overall evaluation of learners' language skills (Communicative Competence) as a general language proficiency, rather than attempting to determine this with a battery of skills' tests.

In this study, the three C-Tests, which are in use at the centre for Professional and Continuing Education (PaCE), Massey University, New Zealand, and which have been tested with international ESL cohorts in New Zealand, are evaluated for a mono-cultural Vietnamese cohort. The study aims to investigate or confirm whether these same C-Tests can be used to evaluate the overall English proficiency of the Vietnamese students validly, reliably and efficiently. The investigation of these C-Tests will hopefully produce similar results to European, Japanese, Middle Eastern, and New Zealand samples.

2 C-Test as an Assessment Instrument

A quick review of the literature for C-Tests demonstrates the depth and breadth of use that this type of assessment instrument holds (Tabatabaei & Shakerin, 2013; Wilmes, 2007). Scoring the test is simple, fast and can be accomplished with a stencil of selected items, if Classical Test Theory (CTT), item analysis (IDI), and/or reliability checks, demonstrate better discrimination among some items/word choices than others. Poor discriminating items may simply be deleted from the scoring stencil if needed.

2.1 Literature informing this study

Modification of strict C-Test formats and content, based on analysis can enhance criterion-referencing and content validity (Bachman, 2004; Baker, 1997; Davidson & Lynch, 2002; Wilmes, 2007). These tests are an invaluable tool that is easily created, and quickly analysed for reliability and validity using CTT item analysis as a basis (Ainol Madziah, & Noor Lide, 2006; Alderson, Clapham & Steel, 1997; Eckes & Grojahn, 2006; Ikeguchi, 1998b). They are so well conceived that they demonstrate at least fair or moderate (highly significant) relationships among all major English language skills, i.e. reading, listening, speaking, and writing (Davies, 2001; Hiser, 2005, 2012; Lei, 2008; Sigott, 2004). In fact, Herriman (2004) calls the construct underlying C-Tests: general communicative competence (GCC) which is similar to Alderson's (2002) concept, or Farhady and Jamali's (2006) general language proficiency (GLP). C-Tests have also demonstrated measurement of progress in English language studies on short-term intensive programmes (Daller & Phalen, 2006).

Ikeguchi (1998a) reports some reliability differences among four C-Test passage types (genre) in a study by Mochizuki (1994) in Japan with tertiary EFL students. Apparently narrative passages were found to provide the best reliability and concurrent validity. The three tests used in this study are in narrative style for that reason. On the other hand, Mochizuki's indication of *long* narrative passages were counter-indicated by Ikeguchi (1998a) and Farhady and Jamali (2006) whose studies showed better test results with combined scores for several short passages.

The use of C-Tests is generally accepted as reliable and valid with a minimum of work in the area of item analysis and reliability (Chapelle, 1994; Fulcher, 1997; Hiser, 2002, 2010, 2012; Ikeguchi, 1998a/b; Lei, 2008). As stated above, the overall consensus is that C-Tests produce results not only in the area of an individual skill but in the general evaluation of language proficiency (Eckes & Grojahn, 2006; Hastings, 2002; Herriman, 2004; Sigott, 2004; Tabatabaei & Shakerin, 2013). According to Dornyei and Katona (1993), "this language-testing instrument has gained high popularity because of its high reliability, sufficient validity, and remarkable practicality" (p. 35). C-Tests prove to have high criterion (concurrent) and construct validity by demonstrating measurement of the same underlying constructs as the MTELP test (Rouhani, 2008), the Michigan University ESL/EFL tests (Hiser, Ishihara, & Okada, 2003), the TOEIC (Daller & Phelan, 2006; Dornyei & Katona, 1993; Herriman, 2004; Hiser, 2005; Rahimi & Saadat, 2005), the IELTS (Cambridge ESOL, 2010; Hiser, 2010, 2012; IELTS, 2003), and the Japanese STEP-Eiken exam (Ikeguchi, 1998b). Correlations among C-Tests and GLP/GCC sub-skills on several standardized tests also contribute to the high acceptance of validity for C-Tests (Daller & Phelan, 2006; Huhta, 1996).

Because C-Tests measure a unified construct (proficiency) rather than individual subcomponents such as listening, speaking, reading, vocabulary, structure, or writing, there is an opportunity for smaller tertiary programmes to place students at correct levels with a minimum of testing, evaluation, or expense by using such an instrument. Kim's (2006) study on designing a skills test for speaking illustrates the type of results in testing that demonstrates unified constructs. She was not able to break down what she assumed would be the four components of communicative competence within the single skill of speaking (Kim, 2004), and firmly established the unity of a speaking ability paradigm. But the C-Test does correlate well with other language skills show-

ing its overall inclusion of individual components in evaluating general proficiency. The C-Tests when used along with specific four skills tests provide a convenient fifth score, an overall score, a tie-breaking score in assessing individual levels of proficiency. The additional skill test scores provide detailed insight into the structure of individuals' language ability for class grouping if larger programmes can provide for more facilities and faculty for additional classes.

2.2 Important issues in creating a C-Test

A C-Test is created by cutting alternating words in a text into halves, and then asking students to restore the words in the passage (Grotjahn, 1987). Some words have an odd number of letters, in which case the shorter number of initial letters is used to create the test—the greater number of letters is cut from the end of the word if an odd number of letters exists. An introductory sentence is left intact along with a closing sentence at the end. In creating a C-Test, there are crucial format issues to follow, such as underlining the first part of the words for the test item at the beginning of enough blank space for writing the remainder of the word. When the first 'half' of the word is not underlined, some students write the entire word in the blank making scoring a bit problematic. Instructions for the test need to be completely clear to the students as they can provide some clues to as to whether their choice of completed word is accurate or not. If the answer contains too many or too few additional letters, the student can see that it is obviously incorrect. See a sample C-Test passage and a set of instructions often used on C-Tests in Appendix A.

The main alternative to a genuine C-Test is a similar test form where an arbitrary (but fixed) number of words are selected to mutilate (Jafarpour, 1995). Fixing and following an alternative number of *nth* words to cut is an acceptable procedure and works well in discriminating between abilities. Care should also be taken not to eliminate an item/word because it seems too easy or too difficult. Not only is this assumption almost always wrong, it destroys the range of students that the test can evaluate. Logically, if all the easy words are taken out of the test, it will only be accurately evaluating the better students—the range of scores declines. If the "hard" words are taken out of the bank of items, then the test will only evaluate the lower level of proficiency—again, there will not be a full range of scores to consider. When the test seems organised, formatted and ready to use, native speakers should attempt the passage completion. This piloting will point out awkward, confusing, or alternative answers that may be possible but not noticed by the test writer. Perhaps some sentences need to be re-worded or eliminated, or 2–3 possible answers accepted as correct in marking/scoring. This is where stencil marking becomes quite efficient.

Success on C-Tests requires not only appropriate spelling, pronunciation, grammar, vocabulary, and schema knowledge, but also comprehension and cohesion. None of these is evaluated independently on separate scales. C-Tests are holistic instruments. Pronunciation, for example, only comes to the attention of examiners when the misspelling of otherwise correct items shows confused use of what is known as *minimal pairs discrimination.* This may occur in cases where B is confused with V in the completed word lo ≠ lober [lover], or in the L/R confusion which lLeads to the spelling spilal for spiral. Markers of C-Tests may also notice the misuse of B and P, or D and TH (particularly with Arab students) in spelling; or L for N in Chinese test takers. These obviously are not spelling mistakes when they appear in seemingly 'correct' attempts at completing words, but they do contribute to correct communicative performance (Communicative Competence) in English, and must be marked as incorrect in scoring. This pronunciation issue is an aspect of proficiency measured by the C-Test as mentioned above, but the redundancy of English, schema knowledge, cohesion, and context should provide enough clues to establish correct choices (including spelling) for completion of the mutilated words which is probably why C-Tests are based on completing text passages rather than isolated, individual sentences.

2.3 Context of this study

The reliability scores for the study in Hiser (2005) demonstrated acceptable Cronbach's alpha scores of 0.6998, 0.8820, 0.8945, and an overall value of 0.9468 for the three C-Tests used, when

combined (N=99). Correlations among the seven sub-skill factors there ranged from 0.467 for a speaking test and C-Test.1, to 0.816 between the combined C-Test scores and the students' total scores on the TOEIC test (all highly significant). The Scree Plot when all variables for that study were factored produced one main factor (GLP/GCC), just as Herriman's (2004) study did.

Previous research has also shown moderate to strong correlations for these C-Tests with each of the four basic language skills mentioned above (Hiser, 2002, 2005). The strongest correlations are usually with the overall values established for an assessment rather than with any particular skill, but the strongest Pearson's correlation coefficient (r) values among skills and C-Tests are usually found with writing skills (Lei, 2008). Reading might be another area where good relationships are intuitively expected and yet these often may not be so strong. The weakest correlations are found with speaking ability, but the amazing point is that there is *any* relationship at all with oral production skills (Shohamy, 1982). This is another fact that points to the measurement of general proficiency (GLP/GCC) rather than independent skills' evaluation by the C-Test.

The Massey University's centre for Professional and Continuing Education (PaCE) offers five levels of placement both in ESOL and in a programme for direct university entrance (DEEP) in lieu of IELTS scores. C-Tests are used as part of placement for both. These tests provide comprehensive proficiency evaluations to supplement other more specific English language skills such as reading ability or vocabulary knowledge, tested at entrance (Hiser, 2010). The C-Test used on the PaCE Placement Test—*The United Nations*—is C-Test 1 in this study and has been well analysed for reliability and validity. The C-Tests were administered to the sample in Vietnam not for the purpose of a placement test, as the participants have undergone one or two years of English studies, but as part of their EFL course work. The purpose of this study is, therefore, to investigate whether these particular C-Tests can be used as reliable English proficiency measures with various samples and ethnic groups of international EFL cohorts. The academic status of the sample will support the validity of the tests if they indicate a certain consistency of scores, i.e., a limited range of proficiency.

2.4 Research questions

Possible answers and insights to the following questions may present themselves in investigating results of the C-Tests for the sample of Vietnamese tertiary students.

- How well does the item analysis demonstrate discrimination among test-takers?
- What reliability and validity do these C-Tests offer in comparison to the New Zealand studies with a mixed international sample?
- Is it reliable and valid to use these same C-Tests as international (cross-cultural) indicators of general English proficiency in Vietnam?

3 The current study

3.1 Sample

To be eligible for studying at a university, Vietnamese students need to pass the national university entrance examination given annually. They are then randomly arranged into different classes (not by ability or proficiency). Their achievement in coursework determines their advancement and future placement. The students come from different regions and provinces in the country, including mountainous, remote highlands, rural and urban areas. Most of them have studied English since secondary school except a small number who have studied English less, due to limited EFL education in remote areas. The students are said to be mostly of pre-intermediate level of English at the beginning of their first-year studies of English.

The research sample consisted of 101 participants in the 18-22 age group doing first or second year academic English at a university in Vietnam. They studied in the Bachelor of English Programme, majoring in Business English, which takes four years to be completed. In the first two

years, they mainly study academic English, while in the last two years, they specialise in Business English. The number of female participants outnumbered the number of male counterparts. A gender breakdown of the sample consisted of 18.8% (19/101) male students compared to 81.2% (82/101) female students. There were not enough male participants to allow the analysing of gender differences as Tabatabaei and Shakerin (2013) say would appear.

3.2 Procedure: The Instruments

The three C-Tests used at PaCE have been established as being valid and reliable for the international ESL cohorts (Hiser, 2010, 2012) in New Zealand. (See Appendices B, C, & D for the content, items selected from the test passages, and the scoring rubrics.) These C-Tests were administered at Massey in three 15-minute sessions with careful pre-test instructions as to how to complete the blanks indicating the strategy of counting the initial number of letters so that attempted answers would comply with requirements for the same number of letters N (half the word) or N+1 if necessary for a word with an odd number of letters.

The same procedure with the same test instruments was followed in Vietnam. These three C-Tests were administered to a sample of 101 learners of English at the university in Vietnam with some confidence since they were previously used with mixed ethnic groups in New Zealand (Hiser, 2010, 2012). The procedure for developing a test—demonstrating reliability and validity—can be followed in the write-up of these studies for the three tests used here.

The lecturer in Vietnam provided all the students with information about the C-Tests and strategies for attempting them beforehand. They had a chance to ask questions about the scoring of the tests. For ethical considerations, the students were allowed to choose not to have their test scores included in the study sample—none of them chose that option. Participants of the study spent 45 minutes in three 15-minute sessions attempting to complete the passages. Then the papers were collected for marking. The answer sheets were subsequently forwarded to New Zealand for repeated marking and analysis. Once the Vietnamese papers were scored, the data were then analysed for descriptive statistics, reliability coefficients, and Pearson's correlation coefficients among the C-Tests.

4 Results

A general discussion of points discovered in the Vietnamese set of data will be informative in relation to the original data collected in New Zealand. Since the tests have been validated and shown to be reliable for international cohorts of students in New Zealand, the design of the study does not attempt to develop the instruments as 'new' material to be trialled. It is simply an exploratory study to discover the comparability of results to the New Zealand work.

4.1 Variable description

The first variables to be examined in the data set were the total scores on each C-Test independently. There are 20 items and therefore a possible score of 20 on each test with an overall score for the three (Total 60) being the possible cumulative score for the three tests collectively. Coding on the SPSS spreadsheet was either a '0' for an incorrect item answer or a '1' for a correct student answer. This made total scores and reliability easy to calculate.

Figure 1 presents the histograms for each set of scores including the total for the 60 items attempted. All four sets of scores indicate a slight skew to the right (negative), indicating the tests were somewhat easy for the Vietnamese cohort. The mean scores as well as the minimum score for the range were all considerably above the medians. The mean scores and descriptives for each of the four variables are listed in Table 1, taken from the SPSS spreadsheet. This shows the exact amount of skew or kurtosis visible in the histograms, the variance in the four sets of scores, and the standard deviation for each.

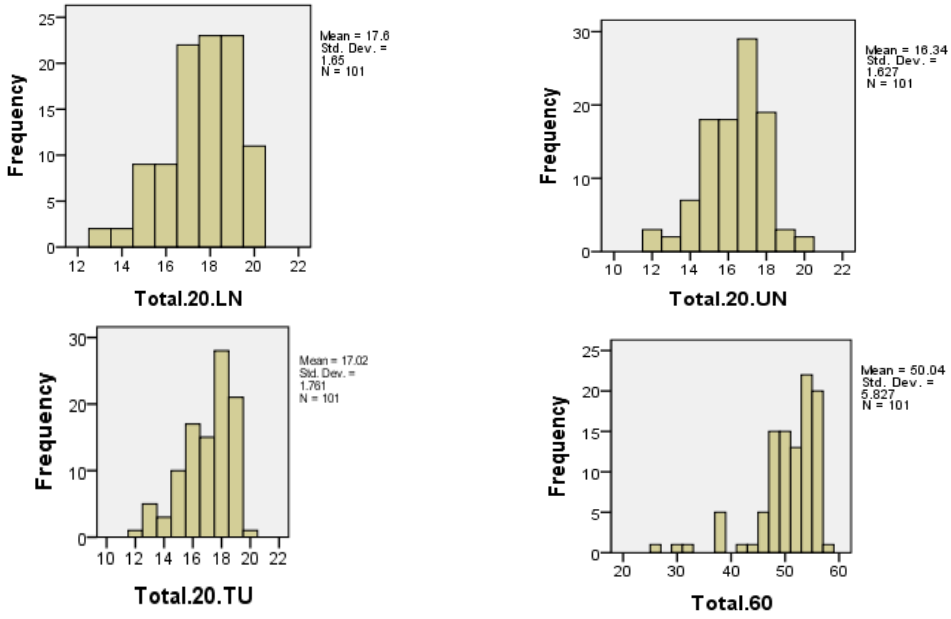

Fig. 1. Histograms for Total Scores on the Vietnamese results, C-Tests 1, 2, 3

Table 1. Descriptive statistics for total scores on the Vietnamese results, C-Tests 1, 2, 3

	N Sample	Min. Score	Max. Score	Mean	Std. Deviation	Variance	Skew	Kurtosis
Total.20.U	101	12	20	16.34	1.627	2.646	-.477	.303
Total.20.L	101	13	20	17.60	1.650	2.722	-.623	.008
Total.20.T	101	12	20	17.02	1.761	3.100	-.782	.005
Total.60	101	26	57	50.04	5.827	33.958	-1.810	3.935

4.2 Item difficulty

Item difficulty is the first item characteristic in CTT to be determined. This is a common practice as tests are often rejected as reliable measures of examinee performance due to the misfit between item difficulty and the examinees' ability (Ainol Madziah & Noor Lide, 2006; Brown, 1996).

There is evidence from the table of frequencies (Appendix F-1, F-2, and F-3) that 16 out of 60 items (26.7%) showed no discrimination between the students, which means all the students did these items correctly. 13 out of 60 items (21.7%) show good discrimination between the students; they include items Q02, Q03, Q10, Q20, L02, L03, L06, L13, L20, T01, T02, T14, T17 which discriminate well (20~80% range of frequency) for both low and high scoring students. The data were divided into two groups for this comparative evaluation. The first group, 'LOW' included scores on the combined item totals of 26~51/60 for 49 cases. The second group, 'HIGH' included cases scoring 52~57/60—a cohort of 52 students. There were 11 items that discriminated well for one or other of the low-high groups but not for both. These items were Q12, Q13, L03, L04, L14, L20, T01, T08, T09, T14, and T16. The remaining items were done by most of the students, except two difficult items, Q13 and T18, which the students answered incorrectly with 80.2% and 94.1% respectively. These are the only two test items that the entire sample found too difficult.

From the results above, the matter of scoring tests should be approached with caution: just because an item is on a test does not mean that it must be marked and included in the score. If test-takers in the sample all (always) get one or two of the words correct or incorrect—too easy or too difficult—it does not mean they have to be scored. Neither does it mean that they should be eliminated or revised. They may be excellent items in expanding the range of scores for the tests. This cohort having been somewhat streamed by admission standards and previous studies, demonstrates the tests' ability to evaluate proficiency by accurately placing the sample at its general level of English skills.

A stencil can be used for marking which only allows scoring of items that show good discrimination. In this case, perhaps 13 out of 60 test items (the ones that show good discrimination between the students) could be marked. This is a quick way to score when only one possible answer (or maybe two) can be correct. The test items with little or no discrimination among the students show that the ability of the sample was higher than the pre-intermediate level. This may have been due to the streaming of students at entrance to the programme at the Vietnamese university, as a minimum English proficiency may have been required for entrance or placement, and that would have limited the range of ability among the sample and cause a skew to higher scores and the kurtosis—an indication that the tests overall were too easy for the cohort.

4.3 Reliability

Reliability is used to measure consistency of answers for students. The accepted value of 0.700 on a test of Cronbach's alpha is generally expected for tests of ability (Cronbach, 1951; Hatcher, 1994; Anastasi, 1997; Brown, 1996). Previous studies with C-Tests have provided reliability scores for a set of three C-Tests used with Japanese ESL students in New Zealand at acceptable levels—0.6998, 0.882, 0.8945, and 0.9468 for the combined scores given on a total of 140 items (Hiser, 2005). In the present study, the individual C-Tests with 20 items each produced low to moderate alpha values for the analyses (Table 2). The combined reliability coefficient for the total 60 items rose to nearly 0.700, a much more acceptable value than that calculated for the individual C-Tests used here. The items which performed the least effectively in this combined score variable were (i) Q11 and Q15 from *The United Nations*, (ii) L05 from the *League of Nations*, and (iii) T05 from *The Tuatara* (see Appendix E-1 to E-4). The worse performing items on individual tests must be combined with items on the other C-Tests to fit into the calculations more successfully leaving these three (out of 60) to lower the alpha value; although the weakness of the items is actually quite small. None of the four items if deleted would raise the Cronbach's alpha value by more than 0.37 for the overall combination of the test values. Probably not of importance, but of interest, is that these four items did fall on each of the individual passages.

Table 2. Reliability coefficients for all C-tests (Vietnamese results)

Title of the C-Test	Cronbach's alpha	No. of items	N/Sample size	Item coding
C-Test 1: The United Nations	0.427	20	101	Q01~20
C-Test 2: The League of Nations	0.396	20	101	L01~20
C-Test 3: The Tuatara	0.491	20	101	T01~20
Combined scores for all C-Tests	0.695	60	101	

A direct comparison of alpha scores for the two administrations of the C-Tests in Table 3 shows the similarity between the two sets of results. While the Vietnamese sample shows lower alpha values, it must be remembered that the range of ability found in the cohort was limited, whereas the New Zealand sample spanned a wider range of proficiency and included cultural differences that may be reflected in broader international schema and comprehension—more sophisticated knowledge of the content of these particular tests. This difference is appropriate not only in communicative language development, but may also be a result of not including higher-level pro-

ficiency students in the present study. Higher-level students would include those with wider, higher reading comprehension, greater schema exposure, and probably more international experience.

Table 3. Comparative reliability between two studies of C-Tests 1, 2, 3

	Vietnamese sample	International Sample (New Zealand)
C-Test 1	0.427	0.6998
C-Test 2	0.396	0.8820
C-Test 3	0.491	0.8945
Combined alpha for the 3	0.695	0.9468

4.4 Validity

Lado (1961) says, "Does the test measure what it claims to measure? If it does, it is valid". This is quite a simple and direct way of defining and determining validity. The meaning of test scores in English language assessment (Chapelle, 2011 p.717) "can refer to a variety of constructs such as knowledge of wh-question formation, reading comprehension, or language ability". (See also Bachman, 2007; Brown, 2005; Chapelle, 1998; McNamara, 1996; Messick, 1989).

Content validity is generally used to assess whether a test is measuring what it is supposed to measure (Dornyei & Katona, 1992; Validity of a Test, n.d.). Face validity is obvious in this study—there are no mathematics or history questions on the C-Tests.

> "A quantitative method of assessing test validity is to examine each test item. This is accomplished by reviewing the discrimination (IDI) of each item. If an item has a discrimination measure of 25 percent or higher, it is said to have validity—it is doing what it is supposed to be doing – discriminating between those that are knowledgeable and those that are not. If an item has a discrimination measure of 25 percent or higher, it is said to have validity". (Validity of a Test, n.d.)

In the present study we allowed a range of items between 20% and 80% discrimination as being acceptable.

Construct validity (Chapelle, 1999, 2011, 2012) can be demonstrated by an internal analysis of the correlations on various components and was assumed, due to the prior investigations in New Zealand. We chose to evaluate the three tests against themselves since we did not have other English scores to assist and the study was conceived as exploratory. Table 4 presents the results of the correlations. Among the three test scores of 20 items each and the collective total of scores (60 items), a Pearson's correlation analysis showed moderate relationships with highly significant coefficients (P = 0.001 two-tailed) for the sample of 101 students. The C-Test score which produced the greatest contribution to the collective score for 60 items was the third C-Test, *The Tuatara* (r = 0.664). All r values were highly significant (P≤ 0.01) for the sample. This is a clear indication that all three C-Tests are measuring the same general construct—English Proficiency—given the face/construct validity of the tests. These results compare favourably with the New Zealand range of correlation results reported above—from 0.467 (for a speaking test, the lowest) to 0.816 (the highest, between the combined C-Test scores and the students' total scores on the TOEIC test, all highly significant). These results are carried over to this study to indicate concurrent validity.

Table 4. Pearson's correlation coefficients among all the C-tests (Vietnam)

		Total.20.U	Total.20.L	Total.20.T.	Total.60
Total.20.U	Pearson correlation	1	.479**	.576**	.617**
Total.20.L	Pearson correlation	.479**	1	.433**	.555**
Total.20.T	Pearson correlation	.567**	.433**	1	.644**
Total.60	Pearson correlation	.617**	.555**	.664**	1

**Correlation is significant at the 0.01 level (2-tailed).

Criterion validity was accepted by the academic position of the cohort at the university in the English language department. The limited range of scores (31) and the appropriate clustering of values in the Gaussian curves for the histograms of each set of C-Test scores, support indications of validity.

5 Conclusions

It can be said that C-Tests are powerful evaluators of English proficiency if well designed, whether the construct is called general language proficiency or general communicative competence. This study has shown that the C-Tests used having been tested with international ESL cohorts in New Zealand, are reliable English proficiency measures with Vietnamese EFL cohorts as the tests offer an acceptable reliability score for the combination of the three C-Tests and significant correlations among them. If the academic position of the students at the university, i.e. their advancement in first and second year work there, is taken as an acceptable confirmation of some English ability, along with the internal correlations among the three C-Tests, then the validity of the tests is also established.

5.1 Implications

The implications which come to mind for test developers of C-Tests need to be discussed. For test items with little or no discrimination among the test-takers, perhaps these items might be excluded—if only in the scoring. Instructors could only mark certain items (not counting them in the scoring) to more accurately reflect ability. But, for the sake of formatting, introducing the test procedure, and reducing test-takers' anxiety, easy items—particularly if they fall at the beginning—and some low discrimination items should be left if not scored. To ensure the validity, the C-tests should be given to a sample that expands the range of ability for an ethnic group.

5.2 Comments on research questions

Research Question 1: How well does the item analysis demonstrate discrimination among test-takers? The results point to a surprising finding of a fairly homogenous sample, when a wider range of ability had been expected. The New Zealand sample had a wider range of proficiency, which was good for test development, but the C-Tests clearly demonstrated their evaluation accuracy in profiling the Vietnamese group.

Research Question 2: What reliability and validity do these C-Tests offer in comparison to the New Zealand studies with a mixed international sample? The tables and charts on Reliability in the Appendices give detailed information on this when compared to the original study results. There are fewer strong items here, but a limited range of proficiency would produce such results if the tests are performing correctly.

Research Question 3: Is it reliable and valid to use these same C-Tests as international (cross-cultural) indicators of general English proficiency in Vietnam? From these exploratory results it appears they will be good evaluators although further/on-going research will most likely demonstrate this in depth.

For the present study, further administrations of the tests to alternative between ability groups in-country should be attempted. Additionally, the same battery of C-Tests should be given to various ethnic groups for comparison and contrast of results (in the DI, for example). The C-Tests as used here seem to deliver the support in assessing students which is needed for accuracy and fairness with a minimum amount of time needed in developing, marking, and analysing. Expansion of the research on the tests for other, and between, ethnic groups should be attempted with the same instruments.

Language testing is often an area where teacher training in language skills and methodology is the weakest (American Federation of Teachers, 1990; Brindley, 2001; Coniam, 2009; Lukin, Bandalos, Eckhout & Mickelson, 2004; Shaffer, 2003; & Taylor, 2009). Many trainees are in the pro-

fession to avoid intense mathematical studies and statistics. C-Tests are, as Dornyei and Katona (1993) say, 'teacher friendly', and basic skills in testing and data analysis can easily be taught giving instructors more confidence in evaluating students fairly. Encouragement and support on the part of management in professional development should be offered to instructors who are in need of testing skills.

This study as an exploratory work has indicated that further research with a wider range of student ability and larger sample would provide expanded generalizability and options for greater comparisons between ethnic groups, but the results are promising.

References

Ainol Madziah, Z., & Noor Lide, A. K. (2006). Classical and Rasch analyses of dichotomously scored reading comprehension test items. *Malaysian Journal of ELT Research, 2,* 1–20.

Alderson, J. C. (2002). Testing proficiency and achievement: principles and practice. In J. A. Coleman, R. Grotjahn, & U. Raatz (Eds.) *University language testing and the C-Test* (pp. 15–30). Bochum: AKS-Verlag.

Alderson, J. C., Clapham, C., & Steel, D. (1997) Metalinguistic knowledge, language aptitude and language proficiency. *Language Teaching Research*, *1*, 93–121.

American Federation of Teachers. (1990). *Standards for teacher competence in educational assessment of students*. Washington, D. C.: Author.

Anastasi, A., (1997). *Psychological Testing*. New York, NY: Prentice Hall.

Bachman, L. F. (2004). *Statistical analyses for language assessment*. Cambridge: Cambridge University Press.

Bachman, L. F. (2007). What is the construct? The dialectic of abilities and contexts in defining constructs in language assessment. In J. Fox, M. Wesche, D. Bayliss, L. Cheng, C. Turner, & C. Doe (Eds.) *Language testing reconsidered* (pp. 41–71). Ottawa: University of Ottawa Press.

Baker, R. (1997). *Classical test theory and item response theory in test analysis*. Lancaster: Centre for Research in Language Education, Lancaster University.

Brindley, G. (2001). Language assessment and professional development. In C. Elder, A. Brown, K. Hill, N. Iwashita, T. Lumley, T. McNamara, & K. O'Loughlin, (Eds.), *Experimenting with uncertainty: Essays in honour of Alan Davies* (pp. 126 – 136). Cambridge: Cambridge University Press.

Brown, J. D. (1996). *Testing in Language Programmes*. Upper Saddle River, NJ: Prentice Hall Regents.

Brown, J. D. (2005). Language test validity. *Testing in language programs: A comprehensive guide to English language assessment* (pp. 220–251). New York, NY: McGraw-Hill.

Cambridge ESOL. (2010). *IELTS information for candidates*. Cambridge: Cambridge University Press. Retrieved from http://www.ielts.org

Chapelle, C. A. (1994). Are C-Tests valid measures for L2 vocabulary research? *Second Language Research*, *10*(2), 157–187.

Chapelle, C. A. (1998). Construct definition and validity inquiry in SLA research. In L. F. Bachman and A. D. Cohen (Eds.). Interfaces between second language acquisition and language testing research (pp. 32–70). Cambridge: Cambridge University Press.

Chapelle, C. A. (1999). Validity in Language Assessment. *Annual Review of Applied Linguistics, 19:*254–272.

Chapelle, C. A. (2011). Validation in Language Assessment. In E. Hinkel (Ed.) *Handbook of Research in Second Language Learning, Vol. 2* (pp. 717–730). Abingdon-on-Thames: Routledge.

Chapelle, C. A. (2012). Validity argument for language assessment: The framework is simple… *Language Testing, 29*: 19–27.

Coniam, D. (2009). Investigating the quality of teacher-produced tests for EFL students and the effects of training in test development principles and practices on improving test quality. *System, 37*, 226–242.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*, 297–334.

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. 101–119). Frankfurt: Peter Lang.

Davidson, F. & Lynch, B. (2002). *Testcraft: A teacher's guide to writing and using language test specifications*. New Haven, CT: Yale University Press.

Davies, A. (2001). The logic of testing languages for specific purposes. *Language Testing*, *18*(2), 133–147.

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

Dornyei, Z. & Katona, L. (1993). The C-Test: A teacher friendly way to test language proficiency. *English Teaching Forum, 31*(1): 34–36.

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

Farhady, H., & Jamali, F. (2006). Varieties of C-Test as measures of general language proficiency. In H. Farhadi (Ed.), *Twenty-five years of living with applied linguistics: A collection of articles* (pp. 287–302). Oxford: B. Blackwell.

Fulcher, G. (1997). An English language placement test: Issues in reliability and validity. *Language Testing*, *14*(2), 113–138.

Grotjahn, R. (1987). How to construct and evaluate a C-Test: A discussion of some problems and some statistical analyses. In R. Grotjahn, C. Klein-Braley, & D. K. Stevenson (Eds.), *Taking their measure: The validity and validation of language tests* (pp. 219–253). Bochum: Brockmeyer.

Hastings, A. (2002). In defense of C-Testing. In R. Grotjahn (Ed.), *Der C-Test. Theoretische grundlagen und praktische anwendungen* [The C-Test: Theoretical foundations and practical applications] (pp.11–25). Bochum: AKS-Verlag.

Hatcher, L. (1994). *A step-by-step approach to using the SAS(R) system for factor analysis and structural equation modelling*. Cary, NC: SAS Institute.

Herriman, M. (2004). Can TOEIC connect reception with production in English? *NUCB Journal of Language, Culture and Communication*, *6*(1), 13–26.

Hiser, E. A. (2002). *Validity of C-Test cloze for tertiary EFL students in Japan*. Proceedings of the JACET Annual Conference, Shizuoka, Japan.

Hiser, E. A. (2005). *Second language assessment, placement, TOEIC, and home grown vegetables.* Paper presented at the ALANZ Symposium 2005: Second Language Assessment and Second Language Learning, Victoria University, Wellington, New Zealand.

Hiser, E. A. (2010). *Mediating placement: Using C-Tests*. Presentation given at CLESOL Dunedin, New Zealand.

Hiser, E. A. (2012). Your programme, your placement, and IELTS: Emerging opportunities in English language evaluation. In G. Skyrme (Ed.), *CLESOL 2012: Proceedings of the 13th National Conference for Community Languages and ESOL*. Retrieved from http://www.tesolanz.org.nz/

Hiser, E., Ishihara, K., & Okada, T. (2003). Modifying C-Test for practical purposes. *Doshisha Studies in Language and Culture*, *5*(4), 539–568.

Huhta, A. (1996). *Validating an EFL C-Test for students of English philology*. Retrieved from http://ltj.sagepub.com/content/23/3/290.refs

IELTS Academic Test. (2013). Retrieved from www.ieltshelpnow.com/ielts_academic_test.html

Ikeguchi, C. (1998a). The four cloze types: To each its own. *Tsukuba Women's University Research Journal*, Tsukuba, Japan.

Ikeguchi, C. (1998b). Do different C-Tests discriminate proficiency levels of EL2 learners? *JALT Testing & Evaluation SIG Newsletter*, *2*(1), 2–10. Retrieved from http://jalt.org/test/ike_1.htm

Jafarpour, A. (1995). Is C-testing superior to Cloze? *RELC Journal*, *30*(2), 86–100.

Kim, H. J. (2004). Task-based performance assessment for teachers: Key issues to consider. *Teachers College, Columbia University Working Papers in TESOL & Applied Linguistics*, *4*(2).

Kim, H. J. (2006). Providing validity evidence for a speaking test using FACETS. *Teachers Columbia University Working Papers in TESOL & Applied Linguistics*, *6*(1), 1–37.

Lado, R. (1961). *Language Testing: The construction and use of foreign language tests.* London: Longman.

Lei, L. (2008). Validation of the C-Test amongst Chinese ESL learners. *The Journal of Asia TEFL, 5*(2), 117–140.

Lukin, L., Bandalos, D., Eckhout, T., & Mickelson, K. (2004). Facilitating the development of assessment literacy. *Educational Measurement: Issues and Practice, 23*: 26–31.

McNamara, T. F. (1996). *Measuring second language performance.* London: Longman.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (pp. 13–103). New York, NY: Macmillan.

Mochizuki, A. (1994). C-Tests: Four kinds of text, their reliability and validity. *JALT Journal*, *16*(1), 41–54.

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.

Rahimi, M., & Saadat, M. (2005). A verbal protocol analysis of a C-Test. *Iranian Journal of Applied Linguistics*, *8*(2), 55–85.

Rouhani, M. (2008). Another look at the C-Test: A Validation Study with Iranian EFL Learners. *The Asian EFL Journal*, *10*(1), 154–180.

Shaffer, D. E. (2003). What works in foreign language teaching. *The Internet TEFL Journal, 44*.

Shohamy, E. (1982). Predicting speaking proficiency from Cloze tests: Theoretical and practical considerations for test substitution. *Applied Linguistics*, *3*, 161–171.
Sigott, G. (2004). *Towards identifying the C-Test construct*. Frankfurt: Peter Lang.
Tabatabaei, O., & Shakerin, S. (2013). The effect of content familiarity and gender on EFL learners' performance on MC cloze test and C-test. *International Journal of English Language Education*, *1*(3).
Taylor, L. (2009). Developing assessment literacy. *Annual Review of Applied Linguistics, 29:* 21–36.
Validity of a Test. (n.d.). Oshkosh, WI: University of Wisconsin. Retrieved from http://www.uwosh.edu/testing/faculty-information
Wilmes, C. (2007). *Validation of a German language placement test based on a modified C-Test procedure*. (Unpublished doctoral dissertation). University of Illinois, Urbana-Champaign.

Appendices

Appendix A

C-Test instructions

This is a simple completion task. Please complete the words with blanks. Half of the word is given for you. So, if there are three letters in front of the blank, you must write either three or four more letters to complete the word correctly. If there are four letters in front of the blank, then the correct word has eight or nine letters, and so on.

Example: There are se____ days in a week and twelve mon___ in a year.
Answer: There are seven__ days in a week and twelve months in a year.

Notice that dividing the word 'seven' means that an additional letter needs to be added to complete the word correctly. Instead of two letters, two 'plus one' may be used. The 'plus one' rule applies to all the blanks—three letters or three plus one, four letters or four plus one. Now try the passage below…

Sample C-Tests*

The kea which Clio was studying would usually leave the area once the sun was in the sky which gave us a break for a few hours. Clio us_____ **(1)** this ti_____ **(2)** to ca_____ **(3)** up o_____ **(4)** thesis wo_____ **(5)**. I wo_____ **(6)** sometimes exp_____ **(7)** the ar_____ **(8)**; there we_____ **(9)** a l_____ **(10)** of rea_____ **(11)** nice wa_____ **(12)** in t_____ **(13)** reserve wh_____ **(14)** provided some_____ **(15)** to ta_____ **(16)** up m_____ **(17)** time. Th_____ **(18)** was t_____ **(19)** Hooker La_____ **(20)** Track, wh_____ **(21)** provided breath_____ **(22)** views o_____ **(23)** mountain to_____ **(24)**, glaciers, gla_____ **(25)** lakes, t_____ **(26)** occasional sn_____ **(27)** leopard a_____ **(28)** even a brief gli_____ **(29)** of resi_____ **(30)** native bi_____ **(31)**. We were all well occupied during the trip.

*This passage was adapted from Forest & Bird » Blog Archive » My life as a keaologist: Mt Cook trip #1 found at http://blog.forestandbird.org.nz/my-life-as-a-keaologist-mt-cook-trip-1/ entered on 3 June 2009.

'Kea Cloze' key

Clio used this time to catch up on thesis work. I would sometimes explore the area; there were a lot of really nice walks in the area which provided something to take up my time. There was the Hooker Lake Track, which provided breathtaking views of mountain tops, glaciers, glacial lakes, the occasional snow leopard and even a brief glimpse of resident native birds. We were all well occupied during the trip.

Appendix B

C-Test 1: United Nations Membership

Under the charter, UN membership is open to all peace-loving states that accept the obligations of the organization. New mem_____ **(Q1)** are admi_____ **(Q2)** by a two-thirds vo_____ **(Q3)** of t_____ **(Q4)** General Asse_____ **(Q5)** on t_____ **(Q6)** recommendation o_____

(Q7) the Secu_____ **(Q8)** Council. Si_____ **(Q9)** 1945, membe_____ **(Q10)** has incr_____ **(Q11)** to th_____ **(Q12)** times t_____ **(Q13)** original num_____ **(Q14)**. These n_____ **(Q15)** members a_____ **(Q16)** mainly fr_____ **(Q17)** African a_____ **(Q18)** Asian coun_____ **(Q19)** which h_____ **(Q20)** been European colonies. The UN had grown to 157 member countries by the end of 1981.

Appendix C

C-Test 2: The League of Nations

The forerunner of the United Nations was the League of Nations. This organ_____ **(L1)** was conc_____ **(L2)** in sim_____ **(L3)** circumstances t_____ **(L4)** the Uni_____ **(L5)** Nations b_____ **(L6)** earlier dur_____ **(L7)** the Fi_____ **(L8)** World W_____ **(L9)**. It w_____ **(L10)** established i_____ **(L11)** 1919 un_____ **(L12)** a tre_____ **(L13)** "to pro_____ **(L14)** international coope_____ **(L15)** and t_____ **(L16)** achieve pe_____ **(L17)** and secu_____ **(L18)**". The Lea **(L19)** ceased i_____ **(L20)** activities after failing to prevent the Second World War. It was a good model for the United Nations which was much more successful.

Appendix D

C-Test 3: The Tuatara

The tuatara is one of New Zealand's reptiles. This little lizard gr_____ **(T1)** to about 60 cm and can li_____ **(T2)** to be 100 ye_____ **(T3)** old. These creatures ha_____ **(T4)** been on o_____ **(T5)** planet much lon_____ **(T6)** than peo_____ **(T7)** have. Tuataras have been aro_____ **(T8)** for 220 mil_____ **(T9)** years, and once shared the ea_____ **(T10)** with the dinosaurs. They a_____ **(T11)** active at ni_____ **(T12)**, and dine on ins_____ **(T13)**, small mammals, and bi_____ **(T14)** eggs. Tuataras are now fo_____ **(T15)** only on a few small isl_____ **(T16)** and in zo_____ **(T17)**. They are not fa_____ **(T18)** nor colourful, but you wi_____ **(T19)** be impressed by th_____ **(T20)** skin, eyes and tail. Take the time to learn more about them; you will be glad you did.

Appendix E-1

Reliability for C-Test 1

The United Nations C-Test Vietnamese Sample
Item-Total Statistics UN C-test 1. Overall Alpha = 0.427 / N=101

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
Q01	15.45	2.750	.239	.398
Q02	15.74	2.333	.219	.384
Q03	16.14	2.321	.259	.367
Q04	15.41	2.944	.000	.428
Q05	15.48	2.412	.493	.327
Q06	15.46	2.850	.062	.425
Q07	15.48	2.752	.149	.409
Q08	15.43	2.887	.078	.423
Q09	15.41	2.944	.000	.428
Q10	16.03	2.409	.197	.392
Q11	15.46	2.970	-.099	.450
Q12	15.53	2.531	.279	.373
Q13	16.21	2.686	.074	.431
Q14	15.44	2.888	.045	.426
Q15	15.43	2.967	-.089	.440
Q16	15.44	2.908	.011	.431
Q17	15.42	2.925	.024	.427
Q18	15.43	2.947	-.048	.436
Q19	15.49	2.872	-.003	.439
Q20	15.89	2.438	.161	.407

Appendix E-2

Reliability for C-Test 2

The League of Nations C-Test Vietnamese Sample
Item-Total Statistics LN C-Test 2, Overall Alpha = 0.396 / N=101

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
L01	16.76	3.699	.161	.380
L02	17.11	3.089	.308	.315
L03	17.00	3.212	.275	.332
L04	16.89	3.493	.159	.371
L05	16.74	3.891	-.092	.408
L06	17.32	2.220	.153	.459
L07	16.72	3.860	.000	.398
L08	16.73	3.815	.089	.393
L09	16.73	3.775	.194	.385
L10	16.72	3.860	.000	.398
L11	16.72	3.860	.000	.398
L12	16.73	3.775	.194	.385
L13	17.06	3.552	.045	.408
L14	16.84	3.429	.269	.348
L15	16.72	3.860	.000	.398
L16	16.75	3.705	.191	.378
L17	16.73	3.815	.089	.393
L18	16.75	3.806	.038	.396
L19	16.73	3.835	.038	.396
L20	16.93	3.338	.237	.348

Appendix E-3

Reliability for C-Test 3

The Tuatara C-Test Vietnamese Sample
Item-Total Statistics Tuatara C-Test 3--Overall Alpha = 0.491 / N=101

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
T01	16.20	2.660	.228	.460
T02	16.04	2.998	.111	.485
T03	15.99	3.130	.000	.493
T04	16.08	2.714	.354	.438
T05	16.13	3.053	-.036	.520
T06	16.01	3.010	.207	.477
T07	16.01	3.110	.001	.496
T08	16.16	2.635	.290	.444
T09	16.11	2.858	.151	.478
T10	16.03	3.069	.033	.495
T11	16.02	3.080	.035	.494
T12	16.02	3.080	.035	.494
T13	16.14	2.761	.203	.466
T14	16.30	2.371	.381	.408
T15	16.05	2.908	.205	.470
T16	16.18	2.588	.307	.438
T17	16.44	2.828	.031	.526
T18	16.93	2.945	.157	.478
T19	15.99	3.130	.000	.493
T20	16.00	3.080	.115	.487

Appendix E-4

Overall reliability

Combined C-Tests (60 items) Vietnamese Sample
Item-Total Statistics for combined C-Tests N=101, Overall Reliability = 0.695

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
T01	50.30	17.323	.290	.684
T02	50.15	18.169	.127	.693
T03	50.10	18.455	.000	.696
T04	50.19	17.287	.454	.679
T05	50.24	18.386	-.018	.701
T06	50.12	18.288	.122	.694
T07	50.12	18.349	.071	.695
T08	50.27	17.431	.279	.685
T09	50.22	17.426	.338	.683
T10	50.14	18.324	.055	.696
T11	50.13	18.478	-.036	.698
T12	50.13	18.478	-.036	.698
T13	50.25	17.745	.192	.690
T14	50.40	16.768	.391	.676
T15	50.16	18.297	.049	.696
T16	50.29	16.753	.479	.672
T17	50.55	17.684	.124	.696
T18	51.04	18.160	.117	.694
T19	50.10	18.455	.000	.696
T20	50.11	18.422	.026	.696
Q01	50.14	17.920	.298	.688

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
Q02	50.44	17.037	.270	.685
Q03	50.84	17.449	.190	.691
Q04	50.10	18.455	.000	.696
Q05	50.17	17.092	.527	.675
Q06	50.15	18.088	.171	.692
Q07	50.17	17.779	.282	.687
Q08	50.12	18.349	.071	.695
Q09	50.10	18.455	.000	.696
Q10	50.72	17.153	.263	.685
Q11	50.15	18.371	.019	.697
Q12	50.23	17.391	.337	.683
Q13	50.90	17.788	.149	.693
Q14	50.13	18.276	.102	.694
Q15	50.12	18.652	-.179	.700
Q16	50.13	18.417	.005	.697
Q17	50.11	18.422	.026	.696
Q18	50.12	18.491	-.046	.697
Q19	50.18	18.189	.082	.695
Q20	50.59	16.891	.318	.681
L01	50.14	17.879	.322	.688
L02	50.49	17.222	.244	.687
L03	50.38	17.006	.335	.680
L04	50.27	17.835	.149	.693
L05	50.12	18.470	-.029	.697
L06	50.70	16.717	.067	.732
L07	50.10	18.455	.000	.696
L08	50.11	18.442	.003	.696
L09	50.11	18.200	.287	.692
L10	50.10	18.455	.000	.696
L11	50.10	18.455	.000	.696
L12	50.11	18.200	.287	.692
L13	50.44	17.259	.245	.687
L14	50.22	17.143	.445	.677
L15	50.10	18.455	.000	.696
L16	50.13	18.215	.143	.693
L17	50.11	18.442	.003	.696
L18	50.13	18.437	-.008	.697
L19	50.11	18.483	-.044	.697
L20	50.31	17.085	.355	.680

Appendix F-1

Item Difficulty Analysis / Frequencies*

For combined C-Test Scores, 60 items: (N=101), Vietnamese Cohort

Item no.	Value	Frequency	Valid percentage	Comments on Items**
Q01	0	4	4.0	**Value 0=student mistake
	1	97	96.0	Value 1=student answer correct
Q02	0	36	35.6	good discriminator
	1	65	64.4	
Q03	0	76	75.2	good discriminator
	1	25	24.8	
Q04	0	0	0	zero discrimination, too easy
	1	101	100.0	
Q05	0	8	7.9	
	1	93	92.1	
Q06	0	5	5.0	
	1	96	95.0	
Q07	0	7	6.9	
	1	94	93.1	
Q08	0	2	2.0	
	1	99	98.0	
Q09	0	0	0	zero discrimination, too easy
	1	101	100.0	
Q10	0	63	62.4	good discriminator
	1	38	37.6	
Q11	0	5	5.0	
	1	96	95.0	
Q12	0	13	12.9	
	1	88	87.1	
Q13	0	81	80.2	weak discrimination, difficult item
	1	20	19.8	
Q14	0	3	3.0	
	1	98	97.0	
Q15	0	2	2.0	
	1	99	98.0	
Q16	0	3	3.0	
	1	98	97.0	
Q17	0	1	1.0	probably no discrimination, too easy
	1	100	99.0	
Q18	0	2	2.0	
	1	99	98.0	
Q19	0	8	7.9	
	1	93	92.1	
Q20	0	49	48.5	good discriminator
	1	52	51.5	

Appendix F-2

Item Analysis / Frequencies*

For combined C-Test Scores, 60 items: (N=101), Vietnamese Cohort

Item no.	Value	Frequency	Valid percentage	Comments on Items**
L01	0	4	04.0	**Value 0=student mistake
	1	97	96.0	Value 1=student answer correct
L02	0	39	39.0	good discriminator
	1	62	61.0	
L03	0	28	27.2	good discriminator
	1	73	72.3	
L04	0	17	16.8	
	1	83	82.2	
L05	0	2	2.0	too easy
	1	99	98.0	
L06	0	69	68.3	good discriminator
	1	31	30.7	
L07	0	0	0	non-discriminating, too easy
	1	101	100.0	
L08	0	1	1.0	non-discriminating, too easy
	1	100	99.0	
L09	0	1	1.0	non-discriminating, too easy
	1	100	100.0	
L10	0	0	0.0	non-discriminating
	1	101	100.0	
L11	0	0	0.0	non-discriminating
	1	101	100.0	
L12	0	1	1.0	non-discriminating, too easy
	1	100	100.0	
L13	0	34	33.7	good discriminator
	1	67	66.3	
L14	0	12	11.9	
	1	89	88.1	
L15	0	0	0.0	non-discriminating
	1	101	100.0	
L16	0	3	3.0	
	1	98	97.0	
L17	0	1	1.0	non-discriminating
	1	100	99.0	
L18	0	3	3.0	
	1	98	97.0	
L19	0	1	1.0	non-discriminating
	1	100	99.0	
L20	0	21	20.8	good discriminator
	1	80	79.2	

Appendix F-3

Item Analysis / Frequencies*

For C-Test: The Tuatara, (N=101), Vietnamese Cohort

Item no.	Value	Frequency	Valid percentage	Comments on Items**
T01	0	21	20.8	good discriminator
	1	80	79.2	
T02	0	21	20.8	good discriminator
	1	80	79.2	
T03	0	0	0.0	non-discriminating
	1	101	100.0	
T04	0	9	8.9	
	1	92	91.1	
T05	0	14	13.9	
	1	87	86.1	
T06	0	2	2.0	
	1	99	98.0	
T07	0	2	2.0	
	1	99	98.0	
T08	0	17	16.8	
	1	84	83.2	
T09	0	12	11.9	
	1	89	88.1	
T10	0	4	4.0	
	1	97	96.0	
T11	0	3	3.0	
	1	98	97.0	
T12	0	3	3.0	
	1	98	97.0	
T13	0	15	14.9	
	1	86	85.1	
T14	0	31	30.7	good discriminator
	1	70	69.3	
T15	0	6	5.9	
	1	95	94.1	
T16	0	19	18.8	
	1	82	81.2	
T17	0	45	44.6	good discriminator
	1	56	55.4	
T18	0	95	94.1	difficult item
	1	6	5.9	
T19	0	0	0	non-discriminating
	1	101	100	
T20	0	1	1.0	non-discriminating
	1	100	99.0	

*The full IDI was not calculated as the frequency of each item is a clear indicator of the proportion of the sample that answered the question correctly or incorrectly—a direct reflection of item difficulty in this case.

The items marked 'Q'—are taken from the test *The United Nations,* the items marked 'L'—are taken from the test *League of Nations,* and the items marked 'T'—are taken from the test *Tuatara.*

**Value 0 = student mistake, Value 1 = student answer correct.

*** Items that might prove discriminators in a larger sample with a greater range of ability (beginning to discriminate). Proximity to a 20% / 80% split was used to evaluate difficulty/facility for each item.