

DEVELOPING COMMUNICATIVE LANGUAGE TESTS FOR SENIOR HIGH SCHOOL

Y. M. Harsono

Universitas Katolik Atma Jaya, Jakarta

Abstract: The Communicative Approach of teaching English in senior high school in Indonesia has been adopted since the implementation of The 1984 Curriculum, but the tests—the communicative language tests—(CL Tests) have not been developed and used properly. The objective of the study is to develop CL Tests for senior high school. The procedure of conducting the study consists of three major steps, that is, identifying the objectives, developing the test specification, and developing the CL Tests. The development of the CL Tests in detail consists of fifteen sub-steps from determining what language skills tested, selecting the suitable source materials, up to rewriting the CL Tests to be used as CL Tests alternative for senior high school. The results of the test development reveal that there are fifteen CL Tests consisting of three tests of listening, three reading, three speaking, and three writing tests. The whole tests have construct and content validity, no complete evidence of concurrent validity with report marks and semester test scores, high to very high inter-rater reliability, and no complete practicality.

Key words: developing tests, communicative language tests

Communicative Language Teaching (CLT) in Indonesia has been adopted since 1984 in the 1984 English Curriculum, the 1994 English Curriculum (Depdikbud, 1995), and the 2004 Competency-Based Curriculum (Depdiknas, 2001). The objective of CLT is to develop communicative competence (CC), *i. e.*, the ability to communicate through language. The term CC was first coined by Dell Hymes (1972) as a reaction against the concept of language competence which was proposed by Chomsky (1965). According to Chomsky competence simply means ‘knowledge of the language system’ or the grammatical knowledge. Chomsky’s view of language competence was more psycholinguistic in nature,

whereas Hymes' concept was more sociolinguistic. Hymes' concept of competence includes concepts of appropriateness and acceptability. Hymes' term of CC was interpreted and developed in sociolinguistic area by many sociolinguists. CC consists of grammatical competence, strategic competence, and sociocultural competence (Canale and Swain, 1980a, 1980b), and discourse competence as well (Canale, 1983), and linguistic competence instead of grammatical competence and actional competence as well (Celce Murcia, 1995).

Grammatical competence refers to Chomsky's linguistic competence. It deals with the mastery of verbal and non-verbal linguistic codes. This competence includes the mastery of vocabulary, words, phrases, sentences, and pronunciation. This competence is very much needed to interpret and convey the literal meaning of utterances.

Sociolinguistic competence relates to the ability to select and use suitable linguistic forms for a certain context of communication. This competence also has to do with socio-cultural and discourse rules. Therefore, this sociolinguistic context will depend much on certain factors like the objective of the interaction, the status of the speaker and the listener, and the norms and rules of the interaction.

Discourse competence is related to the mastery or ability of the learners to combine grammatical forms and meaning to form a complete spoken or written discourse or text. The unity of a text or discourse is realized through the use of form cohesion and meaning coherence. Cohesion is the relationship between utterances and grammatical structure devices to help one to interpret the meaning of a discourse. Coherence is the relationship among several meanings in a text.

Strategic competence is the ability to use verbal and non-verbal communication strategies in order to make-up for the weaknesses in communication because of the limitation of the language mastery. In addition, this strategic competence is also used to strengthen the effectiveness of communication.

Celce-Murcia et al. (1995) proposed more detailed model of CC consisting of five components: discourse competence, linguistic competence, actional competence, sociocultural competence, and strategic competence. They divide sociocultural competence of Canale and Swain's model into three: sociocultural, actional, and discourse competence. The terminological change of grammatical competence to become linguistic competence is to indicate that this competence also includes lexis and phonology in addition to morphology and syntax. Actional competence is defined as "competence in conveying and understanding communicative intent, that is, matching actional intent with linguistic form

based on the knowledge of an inventory of verbal schemata that carry illocutionary force (speech acts and speech acts sets)" (Celce-Murcia, 1995:17).

As the consequence of CLT, the tests have to be communicative as well. In communicative language tests (CL Tests), a test has to measure the CC realized in the four language skills. In practice, however, up to now there is no adequate, clear understanding and sample of CL Tests which can be used in Senior High Schools (*SMU*). The problem is how CL Tests are to be developed. The objective of the study is to develop the CL Tests for *SMU* in the four language skills of listening, reading, speaking, and writing each of which is led to other skill to make the test more integrative in manner. With this objective, the CL Tests produced are expected to be used as alternative sets of CL Tests for *SMU* in Indonesia.

METHODS

The procedure used to develop the CL tests is adapted from the model developed by Carroll (1980), Carroll and Hall (1985), and Weir (1990), which consists of four major steps, the first two of which are used to develop the CL Tests: designing the tests, developing the tests, operating the tests, and monitoring the test administration.

The real steps implemented in this development include: (1) identifying objectives, (2) developing the test specification, (3) developing the CL Tests drafts which consists of ten minor steps: (a) determining what language skills tested, (b) selecting suitable source materials, (c) determining the suitable test formats, (d) determining the coverage of the tests and the number of test items, (e) writing the test items, (f) writing the scoring guide and grading, (g) submitting the test materials for critical examination by experienced teachers, testers, and subject experts, (h) studying the comments, suggestions, and opinions of the teachers, testers, and subject experts, (i) rewriting the CL test drafts, and (j) administering the preliminary CL test drafts, (4) analyzing the results of the preliminary tryout of the CL test drafts, (5) revising the CL test drafts, (6) trying out the CL test drafts, (7) analyzing the results of the tryout of the CL tests, (8) revising the CL test drafts for *SMU*.

The subjects for the tryout were third year students of *SMU Negeri 1 Malang* in the 2003/2004 academic year. The administration of the tryout was in October and November 2003. The classes used for the tryout were science (*IPA*), language, and mixed class of science and social who took extracurricular

activity.

The instruments used were questionnaires for experienced teachers and experts to judge or give comments, opinions, and criticisms on the CL tests being developed and the questionnaires for the students doing the CL tests.

The analysis of the CL tests tryout included that of the validity, reliability, and practicality of the tests. The analysis of the validity covered content and concurrent validity with the students' report marks and the students' semester test marks. The analysis of reliability was the inter-rater reliability between rater 1 and rater 2. The analysis of the practicality considered the questions of economy, ease of administration, scoring and interpretation of results.

FINDINGS

This section presents the findings from the analysis of CL tests tryout about the CL tests being developed. The analysis of the CL tests tryout consists of the item quality analysis (whether an item is valid for purpose and content), the validity (concurrent), the reliability (inter-rater), and practicality. The content validity which is used to check whether an item is valid for purpose and content is analyzed based on the syllabus and the test itself. The concurrent validity is analyzed based on the scores of individual students of each CL test, the students' semester test marks, and the students' report marks in that semester. The inter-rater reliability is based on the scores of each CL test given by two raters working independently. The practicality is determined based on the questions of economy, ease of administration, scoring and interpretation (Weir, 1990) and the available and required resources (human, material, and time), resources to design, develop and administer the tests (Bachman and Palmer, 2000). The analysis of CL tests tryout that include the analysis of listening, reading, speaking, and writing tests are presented descriptively based on statistic calculation.

This sub-section describes the analysis of the CL tests tryout results of Listening Comprehension tests 1-3, Reading Comprehension tests 1-3, Speaking tests 1-3, and Writing tests 1-6.

Listening Comprehension Tests Analysis

This analysis includes item quality analysis (whether an item is valid for purpose and content), concurrent validity, inter-rater reliability, and practicality.

Listening Comprehension Test 1: Short Answer Question

In the purpose of the test for *SMU*, it is mentioned that students are expected to understand information from a dialog by identifying what the dialog is about, where it takes place, who the people involved are, and what the purpose of the dialog is. The test itself is about a dialog which is about arts, a topic that is expected to be taught to the students. So from the point of view of the content the test is in line with what is expected in the purpose of the test.

The concurrent validity between the students' scores of the Listening Comprehension test 1 and the students' marks from the most recent report (*Nilai Raport*) shows that the correlation coefficient using Pearson Product Moment was $r = .5188$, $p = .024$, and using Kendall correlation, the coefficient was $r = .4477$, $p = .023$.

The concurrent validity between the scores of the individual students taking the Listening Comprehension test 1 and the students' marks from the semester test of English indicates no significant correlation (the correlation coefficient using Pearson Product Moment was $r = .4219$, $p = .059$, and using Kendall correlation the coefficient was $r = .2667$, $p = .094$).

The inter-rater reliability agreement coefficient between rater 1 and rater 2 of the Listening Comprehension test 1 was very high, $r = .9142$.

From practicality point of view, the teacher does not take much time to prepare short answer questions and to score the students' answers as well. The students just write down the answers briefly and can use their own words.

Listening Comprehension Test 2: Completion

The purpose of the test for *SMU* states that students are expected to understand information from a dialog by identifying what the dialog is about, where it takes place, who the people involved are, and what the purpose of the dialog is. The test itself is about a dialog which is about arts, a topic that is expected to be taught to the students. So from the point of view of the content the test is in line with what is expected in the purpose of the test.

The concurrent validity between the individual scores of the students taking the test of the Listening Comprehension test 2 and the students' marks from the most recent report (*Nilai Raport*) shows that there was no significant correlation (using Pearson Product Moment was $r = .2871$, $p = .150$, and using Kendall correlation the coefficient was $r = .2466$, $p = .130$).

The concurrent validity between the individual scores of the students taking.

The test of the Listening Comprehension Test 2 and the students' marks from the English semester test shows that there was no significant correlation either (using Pearson Product Moment the coefficient was $r = .2192$, $p = .216$, and using Kendall correlation the coefficient was $r = .2090$, $p = .146$).

The inter-rater reliability agreement coefficient between rater 1 and rater 2 of the Listening Comprehension Test 2 was very high, $r = .9756$.

From practicality point of view, preparing completion questions and scoring the students' answers, the teacher does not take much time. In doing the test, the students just write down the answers in a word or phrase from what they have listened.

Listening Comprehension Test 3: Partial Dictation

In the purpose of the test for *SMU*, it is mentioned that students are expected to "understand information from a description of an issue related to themes read by the teacher or fellow student". The test itself is about an issue entitled "Questioning PLN's Choice" which is about economics, a topic that is expected to be taught to the students. So from the point of view of the content the test is in line with what is expected in the purpose of the test.

The concurrent validity between the individual scores of the students taking the test of the Listening Comprehension Test 3 and the students' marks from the most recent report (*Nilai Raport*) shows that there was low significant correlation (using Pearson Product Moment was $r = .4402$, $p = .050$, and using Kendall correlation the coefficient was $r = .3798$, $p = .039$).

The concurrent validity between the individual scores of the students taking the Listening Comprehension Test 3 and the students' marks from the English semester test marks shows that there was no correlation (using Pearson Product Moment the coefficient was $r = .3998$, $p = .126$, and using Kendall correlation the coefficient was $r = .3024$, $p = .119$).

The inter-rater reliability agreement coefficient between rater 1 and rater 2 of the Listening Comprehension test 3 was very high, $r = .9867$.

From practicality point of view, the teacher can prepare and score the students' answers of partial dictation quite easily and quickly. The students are just expected to comprehend what they have heard and then write down the words.

Reading Comprehension Tests Analysis

This analysis includes item quality analysis (whether an item is valid for purpose and content), concurrent validity, inter-rater reliability, and practicality.

Reading Comprehension Test 1: Short Answer Questions

In the purpose of the test for *SMU*, the students are expected to understand information from a reading text and therefore they are expected to be able to answer the questions given about the reading text. The test itself is about non-oil exports which is about economics, a topic that is expected to be taught to the students. So from the point of view of the content the test is in line with what is expected in the purpose of the test.

The concurrent validity between the individual scores of the students taking the Reading Comprehension Test 1 and the students' marks from the most recent report (*Nilai Raport*) shows no correlation using Pearson Product Moment was $r = .2826$, $p = .154$, and using Kendall correlation the coefficient was $r = .2518$, $p = .124$.

The concurrent validity between the individual scores of the students taking the Reading Comprehension Test 1 and the students' marks from the English semester test indicates no correlation. The correlation coefficient using Pearson Product Moment was $r = .3580$, $p = .095$, and using Kendall correlation the coefficient was $r = .3015$, $p = .066$.

The inter-rater reliability agreement coefficient between rater 1 and rater 2 of the Reading Comprehension Test 1 was very high, $r = .9133$.

From practicality point of view, short answer questions, for the teacher, does not take much time to prepare the test and to score the students answers as well. The students can find their own answers from what they have read and answer briefly using their own words.

Reading Comprehension Test 2: Completion

In the purpose of the test for *SMU*, it is mentioned that students are expected to understand information from a reading text and therefore they are expected to be able to scan to find what the text is about and to skim for a particular information from the text. The test itself is about Jazz Going Down which is about arts, a topic that is expected to be taught to the students. So from the point

of view of the content the test is in line with what is expected in the purpose of the test.

The concurrent validity between the individual scores of the students taking the Reading Comprehension Test 2 and the students' marks from the most recent report (*Nilai Raport*) shows no correlation. The correlation coefficient using Pearson Product Moment was $r = .2676$, $p = .167$, and using Kendall correlation the coefficient was $r = .1386$, $p = .268$.

The concurrent validity between the scores of the students taking the Reading Comprehension Test 2 and the students' marks from the English semester test shows no correlation. The correlation coefficient using Pearson Product Moment was $r = .1427$, $p = .306$, and using Kendall correlation the coefficient was $r = .0521$, $p = .399$.

The inter-rater reliability agreement coefficient between rater 1 and rater 2 of the Reading Comprehension Test 2 was very high, $r = .9029$.

From practicality point of view, completion test, for the teacher, does not take much time to prepare and to score the students' answers as well. The students find their own answers from what they have read and write the answers by completing sentences.

Reading Comprehension Test 3: Essay Questions

The purpose of the test for *SMU* mentions that students are expected to understand information from a reading text and are expected to be able to answer questions about the text. The test itself is about "17 TKIs Died in Nunukan Camps: Official" which is about Social and Population Issues, a topic that is expected to be taught to the students. So from the point of view of the content the test is in line with what is expected in the purpose of the test.

The concurrent validity between the scores of the students taking the Reading Comprehension Test 3 and the students' marks from the most recent report (*Nilai Raport*) shows no correlation. The correlation coefficient using Pearson Product Moment was $r = .1660$, $p = .277$, and using Kendall correlation the coefficient was $r = .2153$, $p = .169$.

The concurrent validity between the scores of the students taking the Reading Comprehension Test 3 and the students' marks from the English semester test indicates no correlation either. The correlation coefficient using Pearson Product Moment was $r = .2852$, $p = .151$, and using Kendall correlation the coefficient was $r = .2538$, $p = .104$.

The inter-rater reliability agreement coefficient between rater 1 and rater 2 of the Reading Comprehension Test 3 is very high, $r = .9608$.

From practicality points of view, essay question test is easy to prepare and does not take much time but to score the students' answers does. The students may get difficulties in writing their own answers. So this test is impractical in the scoring from the teacher's point of view and in the writing of the answers from the students' point of view.

Speaking Tests Analysis

This analysis includes item quality analysis (whether an item is valid for purpose and content), concurrent validity, inter-rater reliability, and practicality.

Speaking Test 1: Dialog Retelling

The purpose of the test for *SMU* mentions that students are expected to reproduce and make comments about the content of dialogs. The test itself is about getting part time job which is about jobs, a topic that is expected to be taught to the students. So from the point of view of the content the test is in line with what is expected in the purpose of the test.

The concurrent validity between the scores of the students taking the Speaking Test 1 and the students' marks from the most recent report (*Nilai Raport*) indicates no significant correlation. The correlation coefficient using Pearson Product Moment was $r = .9820$, $p = .061$, and using Kendall correlation the coefficient was $r = .8165$, $p = 110$.

The concurrent validity between the individual scores of the students taking the test of the Speaking Test 1 and the students' marks from the English semester test shows no significant correlation either. The correlation coefficient using Pearson Product Moment was $r = .6547$, $p = .273$, and using Kendall correlation the coefficient was $r = .8165$, $p = 110$.

The inter-rater reliability agreement coefficient between rater 1 and rater 2 of the Speaking Test 1 was very high ($r = .9449$).

From practicality point of view, speaking test does take much time for the teacher both to prepare and conduct the test. Each of the students has to understand the dialog and to reproduce the dialog orally using their own words.

Speaking Test 2: Diagram Retelling

It is mentioned in the purpose of the test for *SMU* that students are expected to read and interpret diagrams: tables, graphs, and charts. The test itself is about four different figures/tables which is about population, economics, and tourism, topics which are expected to be taught to the students. So from the point of view of the content the test is in line with what is expected in the purpose of the test.

The concurrent validity between the scores of the students taking the Speaking Test 2 and the students' marks from the most recent report (*Nilai Raport*) indicates no significant correlation. The correlation coefficient using Pearson Product Moment was $r = .9820$, $p = .061$ ($p > 0.05$), and using Kendall correlation the coefficient was $r = 1.000$, $p = .059$ ($p > 0.05$).

The concurrent validity between the scores of the students taking the Speaking Test 2 and the students' marks from the English semester test using Pearson correlation formula indicates significant correlation, but there was no correlation using Kendall formula. The correlation coefficient using Pearson Product Moment was $r = .9991$, $p = .014$ ($p < 0.05$), and using Kendall correlation the coefficient was $r = 1.000$, $p = .059$ ($p > 0.05$).

The inter-rater reliability agreement coefficient between rater 1 and rater 2 of the Speaking Test 2 is very high, $r = .9919$.

From practicality point of view, speaking test 2 is impractical for the teacher, because it does take much time both to prepare and conduct the test. In this test, the students have to understand the charts, table, and graphs and to reproduce/retell them orally using their own words.

Speaking Test 3: Story or Text Retelling

In the purpose of the test for *SMU*, it is mentioned that students are expected to describe main and supporting ideas of reading text related to themes. The test itself is about National Music Day, a topic about arts which is expected to be taught to the students. So from the point of view of the content the test is in line with what is expected in the purpose of the test.

The concurrent validity between the scores of the students taking the Speaking Test 3 and the students' marks from the most recent report (*Nilai Raport*) shows perfect significant correlation using Pearson formula, but not significant using Kendall formula. The correlation coefficient using Pearson

Product Moment was $r = 1.000$, $p = .00$ ($p < 0.05$), and using Kendall correlation the coefficient was $r = 1.000$, $p = 0.79$ ($p > 0.05$).

The concurrent validity between the scores of the students taking the Speaking Test 3 and the students' marks from the English semester test shows that the correlation coefficient using Pearson Product Moment was $r = 1.000$, $p = .079$ ($p > 0.05$), and using Kendall correlation the coefficient was $r = 1.000$, $p = 0.79$ ($p > 0.05$). Both coefficients are not significantly correlated.

The inter-rater reliability agreement coefficient between rater 1 and rater 2 of the Speaking Test 3 was very high, $r = .9820$.

From practicality point of view, this speaking test is very impractical for both the teacher and the students. For the teacher, this test does take much time both to prepare and conduct the test. For the students, it is very impractical for them because they have to understand the reading text, summarize it, and to reproduce it orally using their own words.

Analysis of Writing Tests

This analysis includes item quality analysis (whether an item is valid for purpose and content), concurrent validity, inter-rater reliability, and practicality.

Writing Test 1: Controlled Writing Tasks

In the purpose of the test for *SMU*, it is mentioned that students are expected to be able to write short announcements related to themes. The test itself is about writing announcement of biology make up class, a topic about education which is expected to be taught to the students. So from the point of view of the content the test is in line with what is expected in the purpose of the test.

The concurrent validity between the scores of the students taking the Writing Test 1 and the students' marks from the most recent report (*Nilai Raport*) both using Pearson and Kendall formula shows significant correlation. The correlation coefficient using Pearson Product Moment was $r = .6465$, $p = .005$ ($p < 0.05$), and using Kendall correlation the coefficient was $r = .4808$, $p = .018$ ($p < 0.05$).

In addition, the concurrent validity between the scores of the students taking the Writing Test 1 and the students' marks from the English semester test using Pearson formula indicates low correlation, but no correlation using Kendall formula. The correlation coefficient using Pearson Product Moment was $r =$

.5400, $p = .019$ ($p < 0.05$), and using Kendall correlation the coefficient was $r = .2798$, $p = .090$ ($p > 0.05$).

The inter-rater reliability indicates high agreement coefficient between rater 1 and rater 2 of the Writing Test 1 ($r = .8215$).

From practicality, this writing test 1 is not practical for both the teacher and the students. For the teacher, it does not take much time to prepare but to score the students' writing. From the students side, writing test is impractical because they have to perform their idea in writing in the essay form using their own words.

Writing Test 2: Answering an Advertisement (Letter of Application)

As is stated in the purpose of the test for SMU, it is mentioned that students are expected to be able to write simple formal letter of application. The test itself is about writing a letter of application in response to an advertisement about job vacancy, a topic which is expected to be taught to the students. So from the point of view of the content the test is in line with what is expected in the purpose of the test.

The concurrent validity between the scores of the students taking the Writing Test 2 and the students' marks from the most recent report (*Nilai Raport*) shows no significant correlation. The correlation coefficient using Pearson Product Moment was $r = .3467$, $p = .103$ ($p > 0.05$), and using Kendall correlation the coefficient was $r = .2404$, $p = .160$ ($p > 0.05$).

The concurrent validity between the scores of the students taking the Writing Test 2 and the students' marks from the English semester test indicates low significant correlation. The correlation coefficient using Pearson Product Moment was $r = .4781$, $p = .036$ ($p < 0.05$), and using Kendall correlation the coefficient was $r = .1983$, $p = .184$ ($p > 0.05$).

The inter-rater reliability agreement coefficient between rater 1 and rater 2 of the Writing Test 2 was high ($r = .8311$).

From the practicality view point, this writing test, like the other test of writing is impractical for both the teacher and the students. For the teacher, preparing this test of writing a letter of application does not take much time but to score the students' writing does. From the students side, it is not simple to do this test because they have to understand the advertisement first, then they have to know the format of letter of application, and finally they have to write the letter in the letter format using their own words.

Writing Test 3: Letter of Request

In the purpose of the test for *SMU*, it is mentioned that students are expected to be able to write simple formal letter of request. The test itself is about writing a letter of request asking for information about universities in Sydney and Melbourne to Australian Embassy, a topic which is expected to be taught to the students. So from the point of view of the content the test is in line with what is expected in the purpose of the test.

The concurrent validity between the scores of the students taking the Writing Test 3 and the students' marks from the most recent report (*Nilai Raport*) both using Pearson and Kendall formula indicates no correlation. The correlation coefficient using Pearson Product Moment was $r = .1863$, $p = .253$ ($p > 0.05$), and using Kendall correlation the coefficient was $r = .1562$, $p = .257$ ($p > 0.05$).

Similarly, the concurrent validity between the scores of the students taking the Writing Test 3 and the students' marks from the English semester test both using Pearson and Kendall formula indicates no correlation. The correlation coefficient using Pearson Product Moment was $r = .0936$, $p = .370$ ($p > 0.05$), and using Kendall correlation the coefficient was $r = .0822$, $p = .353$ ($p > 0.05$).

The inter-rater reliability agreement coefficient between rater 1 and rater 2 of the Writing Test 3 was high ($r = .8124$).

From practicality, similar to the writing test of writing a letter of application, preparing the test of writing a letter of request, the teacher does not take much time but to score the students' writing does take a lot of time. The students have to create their own words to express their intention, then they have to know the format of letter of request, and finally they have to write the letter in the letter format using their own words.

Writing Test 4: Filling out Forms

In the purpose of the test for *SMU*, it is mentioned that students are expected to be able to fill in various forms. The test itself is about filling out a form of curriculum vitae, a topic which is expected to be taught to the students. So from the point of view of the content, the test is in line with what is expected in the purpose of the test.

The concurrent validity between the scores of the students taking the Writing Test 4 and the students' marks from the most recent report (*Nilai Raport*)

shows no correlation. The correlation coefficient using Pearson Product Moment was $r = -.0783$, $p = .391$ ($p > 0.05$), and using Kendall correlation the coefficient was $r = -.0877$, $p = .349$ ($p > 0.05$).

Likewise, the concurrent validity between the individual scores of the students taking the Writing Test 4 and the students' marks from the English semester test both using Pearson and Kendall formula indicates no correlation. The correlation coefficient using Pearson Product Moment was $r = -.0861$, $p = .380$ ($p > 0.05$), and using Kendall correlation the coefficient was $r = -.0916$, $p = .334$ ($p > 0.05$).

The inter-rater reliability agreement coefficient between rater 1 and rater 2 of the Writing Test 4 was high ($r = .8528$).

From practicality view, preparing writing test in the form of filling out a form of curriculum vitae does not take much time but to score the students' writing of the completed form is time consuming. It is not difficult for the students to fill in their own data in the curriculum vitae form. They just complete the form by writing what are asked which are true to them.

Writing Test 5: Summarizing

In the purpose of the test for *SMU*, it is mentioned that students are expected to be able to write a summary of a reading text related to themes. The test itself is about summarizing a text entitled "*PDI Perjuangan Wants Simultaneous Elections in 2004*", a topic about politics which is expected to be taught to the students. So from the point of view of the content the test is in line with what is expected in the purpose of the test.

The concurrent validity between the scores of the students taking the Writing Test 5 and the students' marks from the most recent report (*Nilai Raport*) indicates no correlation, the correlation coefficient using Pearson Product Moment was $r = -.0429$, $p = .440$ ($p > 0.05$), and using Kendall correlation the coefficient was $r = -.1062$, $p = .333$ ($p > 0.05$).

Likewise, there was no correlation between the scores of the students taking the Writing Test 5 and the students' marks from the English semester test. The correlation coefficient using Pearson Product Moment was $r = .0351$, $p = .451$ ($p > 0.05$), and using Kendall correlation the coefficient was $r = .0281$, $p = .451$ ($p > 0.05$).

The inter-rater reliability agreement coefficient between rater 1 and rater 2 of the Writing Test 5 was high ($r = .7990$).

From practicality view, writing a summary of a reading text is time consuming for the students to do because they have to read and comprehend the reading text and then summarize the text in sentences. The teacher also takes a lot of time to score the students' summary although it does not take much time to prepare the test.

Writing Test 6: Writing Composition

It is mentioned in the purpose of the test for *SMU* that students are expected to be able to write a short composition related to the themes. The test itself is about writing a short composition about an important event in the students' school surroundings, a topic which is expected to be taught to the students. So from the point of view of the content the test is in line with what is expected in the purpose of the test.

The concurrent validity between the scores of the students taking the Writing Test 6 and the students' marks from the most recent report (*Nilai Raport*) both using Pearson and Kendall formula shows no correlation. The correlation coefficient using Pearson Product Moment was $r = -.0476$, $p = .433$ ($p > 0.05$), and using Kendall correlation the coefficient was $r = -.0959$, $p = .332$ ($p > 0.05$).

Similarly, there was no correlation between the scores of the students taking the Writing Test 6 and the students' marks from the English semester test both using Pearson and Kendall formula. The correlation coefficient using Pearson Product Moment was $r = -.2522$, $p = .182$ ($p > 0.05$), and using Kendall correlation the coefficient was $r = -.2560$, $p = .107$ ($p > 0.05$).

The inter-rater reliability agreement coefficient between rater 1 and rater 2 of the Writing Test 6 was high ($r = .7431$).

From practicality point of view, although it is very easy and does not take a lot of time to prepare the writing test in the form of writing a composition, it really is time consuming to score the students' composition. In addition, the students had to prepare the idea and then organize the idea in the composition in good sentences. So this writing test is clearly impractical for both the students and the teacher.

The results of the analysis revealed that all the CL Tests had content validity and were reliable, high and very high. Only four tests were significantly correlated with the students' report marks, L Test 1, L Test 3, S Test 3, and W Test 1, and only two tests were significantly correlated with the students' semester

test marks, S Test 2, and W Test 1. Only five out of fifteen tests, L Tests 1, 2, and 3 and R Tests 1 and 2 were practical.

DISCUSSION

The evidence for content validity of all the CL Tests developed is due to the fact that each test developed can be logically established by matching the items and the domains to be measured through table of specification in the process of developing the tests.

The absence of significant correlation between the students' scores of the CL Tests and the students' report marks is probably due to the incomplete mark components used to make the report marks.

Similar to the insignificant correlation between the students' scores of the CL Tests and the report marks, most of the CL Tests are not significantly correlated with the semester test marks because the semester test just covered the test of reading, structure, and vocabulary in the multiple choice format.

The high and very high agreement coefficient between rater 1 and rater 2 shows that later on when the test is used to test students, one rater may be enough to score the students' answers and to yield objective results.

The impracticality of most of the CL Tests developed is mostly as the consequence of the characteristic of the CL Tests one of which is the performance of the students reflected in the students' answers to the tasks given in the tests. These impractical tests should not be exclusively refused to be good tests. There should be a way of conducting the tests so that the impracticality of the tests can be overcome.

CONCLUSION AND SUGGESTIONS

Based on the findings above, the CL Tests produced as a whole can be considered to have construct and content validity, no complete evidence of concurrent validity, high to very high inter-rater reliability, and no complete practicality. The presence of construct and content validity is as the result of the match between the items and the domains to be measured which was logically examined and established through a table of specification in the process of developing the tests. The incomplete evidence of concurrent validity with the semester test scores and the report marks is due to the fact that the semester test does not include all the language skills as required in the CL Tests. The results of high and

very high inter-rater reliability agreement indicates that the scoring guides were clear and thus later on in the implementation of the CL Tests one rater would be sufficient. The impracticality of the tests is due to the time needed to conduct the test and to score the students' answers especially if there are many students tested. The impracticality of the tests should not be exclusively used as the reasons of not choosing the CL Tests as a means of evaluating the students communicative competence. The most important test characteristics that should be considered are the validity especially the construct and content validity and the reliability of the tests.

This study is focused on developing communicative language tests for senior high schools in Indonesia. The references are, therefore, the CL Tests theories, the curriculum used for senior high school and the teaching learning activities in the classroom as well as the tests used by the teachers in senior high schools. The suggestions put forward in this study are directed to English teachers, test constructors, and further researchers.

In order to correctly measure the students' communicative competence in the four language skills especially for senior high school students, four suggestions for English teachers and test constructors are given below.

- (1) Before using the tests, senior high school teachers should realize and choose that/those having the best characteristics as expected, such as, having construct and content validity, reliability, and practicality.
- (2) Considering the characteristics of the CL Tests: validity, reliability and practicality, English teachers and test constructors should not just think of the practicality of the CL Tests when using/developing the tests, but, the most important thing is that they should think of the validity as well as the reliability of the test.
- (3) Speaking tests can be administered several days or even weeks before the written tests are administered in the mid- or end-of semester tests so that all the students can be tested thoroughly by the end of the testing period.
- (4) The CL Tests should not be conducted just once at the end of a semester or year, but they should be conducted any time necessary, for example, at the end of every topic of discussion or during the teaching learning activities which is called process assessment. In this process assessment, teachers can prepare a form/ list of students in the class and students' activities or performance being assessed and during the teaching learning process this form

is completed based on each student's performance.

- (5) Test formats should be chosen by considering the characteristics of communicative competence: assessing the dynamic negotiation of meaning, including measures of both written and spoken language, being context specific, and assessing performance that is observable rather than competence.

This study is developing CL Tests for *SMU* on the basis of the Revised 1994 Curriculum. There are other similar areas of study that can be done at the present time. Suggestions are made for researchers who want to develop similar tests. Among others, the following topics are suggested for further study: (1) developing CL Tests for senior high school (*SMU*) on the basis of the Competency Based Curriculum, the 2004 Curriculum, (2) developing CL Tests for Junior High School (*SLTP*), (3) developing CL Tests for Elementary Schools, and (4) developing CL Tests for Specific Purposes using different criteria which are more standardized than using the same criteria used in this research, the semester test marks and the report marks, because the findings show that the CL Tests developed do not have concurrent validity with those criteria.

REFERENCES

- Bachman, L. F. and Palmer, A. S. 2000. *Language Testing in Practice*. Oxford: Oxford University Press.
- Canale, M. 1983. From Communicative Competence to Language Pedagogy. In Jack C. Richards and Richards W. Schmidt (Eds.) *Language and Communication*. London: Longman.
- Canale, M. and Swain, M. 1980a. *Approaches to Communicative Competence*. Singapore: SEAMEO Regional Language Centre.
- Canale, M. and Swain, M. 1980b. Theoretical Bases of Communicative Approaches to Second Language Teaching and Testing. *Applied Linguistics*. I (1): 1-47.
- Carroll, B. J. 1980. *Testing Communicative Performance*. Oxford: Pergamon Press.
- Carroll, B. J. and Hall, P. J. 1985. *Make Your Own Language Tests: A Practical Guide to Writing Language Performance Tests*. Oxford: Pergamon Press.
- Celce-Murcia, Dornyei, M. Z., and Thurrell, S. 1995. Communicative Competence: A Pedagogical Motivated Model with Content Specifications. *Issues in Applied Linguistics*. 6 (2): 5-35.

- Chomsky, N. 1965. *Aspects of the Theory of Syntax*. Cambridge, Mass.: MIT Press.
- Depdikbud. 1995. *Kurikulum SMU 1994: Garis-Garis Besar Program Pengajaran Bidang Studi Bahasa Inggris*. Jakarta.
- Depdiknas. 2001. *Kurikulum Berbasis Kompetensi: Mata Pelajaran Bahasa Inggris Sekolah Menengah Umum*. Jakarta: Pusat Kurikulum Badan Penelitian dan Pengembangan.
- Hymes, D. H. 1972. On Communicative Competence in Brumfit, C. J. & Johnson, K. (Eds.) *The Communicative Approach to Language Teaching*. 1987. (pp. 5-26). Oxford: Oxford University Press.
- Weir, C. J. 1990. *Communicative Language Testing*. New York: Prentice Hall.