English Language & Literature Teaching, Vol. 23, No. 3 Fall 2017

Insights into Test Construction Practices of Native English Speaking Teachers in College Conversation Courses

Mee Hye Kang [*]
(Eulji University)
Daniel James Cuffey [**]
(Eulji University)
Myeong-Hee Seong[***]
(Eulji University)

Mee Hye Kang, Daniel James Cuffey & Myeong-Hee Seong (2017). Insights into test construction practices of native English speaking teachers in college conversation courses. *English Language & Literature Teaching, 23*(3), 69-88.

This study aims to analyze the college English conversation tests of native English teachers to see if the tests complied with program goals. Also, problems with current test construction practices were highlighted and improvements for similar conversation courses were suggested. In order to do this, we examined the test construction practices of seven native English teachers who taught an English conversation course at a university in Gyeonggi province near Seoul. The teachers' mid-term and final tests were collected and they also expressed their views on testing through a survey and interviews. The teachers' tests were analyzed in terms of the types of assessments used, question types, number of test items, language skills targeted, and task types. The results showed that despite having a standardized course goal and using the same textbook, the testing methods utilized by teachers varied widely and teachers typically employed written tests which varied extensively due to non-standardization. However, the teachers believed oral exams should be employed in various ways: a 1:1 interview, individual presentations, group discussions or role plays. The implications of these assessment methods, along with the future direction of this study were discussed.

[Test construction practices/college English conversation courses/native English instructor]

---

[*]The first author, [**] The second author, [***] The corresponding author

I. INTRODUCTION

In recent years, many studies have been conducted to improve general English programs at universities in Korea (H. Chang, 2009; D. Jeong & H. Kim, 2001; J. Joh, 2002; C. Kim, 1988; E. Kim, 2005; O. Kwon, 1995; J. Park, 1997). However, previous studies on college English programs have focused on changes and improvements holistically while little attention has been paid to test construction practices, which have a big impact in classrooms (E. Kim, 2005). According to Bachman and Palmer (1996) and Hughes (1990), classroom assessments provide valuable insights on learning ability, the program accountability and appropriacy, and the efficacy of teaching. This enables teachers to reflect on their instruction to improve the course. In other words, testing is essential because it helps to diagnose the students' strengths and weaknesses and to evaluate their performance at a classroom level. Furthermore, it can aid in the planning of a program at the administrative level.

Additionally, examining the test construction practices for compliance with course goals in general English programs can be used as an important reference to develop university programs (E. Kim, 2005). English conversation is a mandatory subject taught in general English programs to all students at universities in Korea. We suggest that general English programs use a standardized curriculum in terms of the class objectives, learning process, and the classroom assessments. Simply put, there should be consistency throughout the classes by using the same textbook, course syllabus, and assessments. A great number of researchers have pointed out the problems that occur when a majority of teachers run their classes at their own discretion and assess their students subjectively (D. Cho, 1997; C. Kim, 1988; E. Kim, 2005; K. Pyo, 2003).

A lot of research has been conducted on reading test construction practices in general English programs, with little focus on conversation courses (C. Kim, 1988; E. Kim, 2005; B. Lim, 1995; K. Pyo, 2003). Furthermore, there has been extensive research to improve general English programs, but few studies have focused on how tests should change according to program goals. Also, insufficient attention has been paid to the test construction practices of native speaking English teachers (NSETs) in classrooms. With the emphasis on communicative competence (Canale & Swain, 1980) in Korea, universities in Korea have employed a great number of NSETs to teach English classes in college English programs. Several studies have focused on the effects of NSETs in classrooms. For example, H. Lee's (2001) study explored a general evaluation of NSETs in general English programs. Y. Jeong's (2004) research elucidated that NSETs improve students' conversation skills and T. Yang (2011) indicated that students had strong preference for NSETs in learning spoken English. In addition, J. Lim (2006) proved that

students are generally satisfied with NSETs in classrooms, and H. Choi and J. Kang (2009) discovered that it is effective to run English programs with NSETs.

Considering the fact that there are a great number of native English teachers teaching English conversation courses, it is especially important to investigate how they assess their students in conversation classes. Analyzing the test construction practices of NSETs will not only bring insight into the effectiveness of NSETs in college English programs, but also offer aid in planning and administering English programs at universities. Thus, the purpose of the current study is to suggest implications for English conversation courses in general English programs. The research questions were posed as follows:

1. What are the NSETs test construction methods and how do they employ those methods on their tests?
2. What are the teachers' opinions and suggestions for improvement?

II. LITERATURE REVIEW

1. Language Testing Issues

Many ESL researchers agree that testing plays a pivotal role in the language learning process (Bachman, 1989; Chapelle, 1988; Chapelle & Robert, 1986; Heaton, 1975; McNamara, 2000). Testing is a primary device to enhance learning, motivate students, and assess their performance in the target language (Heaton, 1975). This not only benefits language learners, but also advances teaching pedagogy (Bachman, 1989). As a whole, language testing is not just a way to measure how much learning is taking place, but it also exerts a powerful influence on program goals and course objectives (McNamara, 2000).

Scholars and practitioners agree that language testing is an essential part of language learning and teaching, and the best practices have been researched extensively (Alderson, Clapham, & Wall, 1995; Fulcher, 2010). The trend in English assessments has shifted from receptive skills to communicative performance. This signals a move away from indirect, discrete point tests in favor of direct, authentic communication (Bachman, 1990). Although the testing trend in the Korean EFL context has moved from traditional assessments to alternative assessments (E. Kim, 2005), written tests are generally preferred and employed by many test developers and teachers due to a lack of rater training, time consumption, and difficulties in administering (J. Jeon, 2010).

A number of scholars have noted that testing in the Korean EFL context is still pen-and-paper. Many test makers in Korea prefer using written tests to measure language

learning. Students are also more familiar with written tests in general English programs (E. Kim, 2005). However, it has been argued that written tests do not demonstrate learners' skills when performing in actual situations (Canale & Swain, 1980). If tests are made poorly, and fail to measure skills accordingly, they are likely to produce negative washback effects (Hughes, 1990). Essentially, students will not cooperate with the curriculum and course objectives and will only focus on whatever is on the tests (Brown & Hudson, 1998).

Moreover, assessing students in college English programs using non-standardized tests may lead to unfairness and inequity. General English programs are required for all the students at a university, therefore the quality and characteristics of tests play a critical role in general English courses at universities. Ideally, they might have the same textbook, syllabus, and course objectives. Bachman and Palmer (1996) argued that different test methods affect test takers differently. For example, some test takers will do better on oral interviews whereas other test takers might do better on pen-and-paper tests (C. Kim, 1988). As C. Kim (1988) and E. Kim (2005) have argued, problems may arise if students are assessed using different testing methods. Since students are graded on a curve, having a different score for each question, varying questions items and numbers could result in unfairness.

Bachman (1989) states that test makers should develop and use language tests appropriately. To put simply, teachers should measure what they are supposed to measure using optimal methods. When developing tests, test makers should clearly state what they are testing and accurately measure whatever is intended to be measured (Hughes, 1990). Tests need to be based on clearly stated goals, so they should have validity and reliability. Henning (1987) defines validity as the extent to which a test measures the ability or knowledge that it is purported to measure. He defines reliability as "the consistency of the scores obtainable from a test. It is usually an estimate on a scale of zero to one of the likelihood that the test would rank testees in the same order from one administration to another proximate one." Therefore, teachers should be concerned if their tests give consistent results and measure students appropriately.

2. Aspects of Test Construction Practices in Previous Studies

To date, several studies have focused on test construction in general English programs at universities. B. Lim (1995) has analyzed the test papers of ten secondary schools and administered a survey to 80 teachers. The study revealed that the test construction consisted of discrete-point questions such as grammar, language use, pronunciation, and translation. The teacher survey showed that the teachers did not employ their test papers

according to the rules of the assessments. In addition, teachers noted that they marked the tests subjectively without sharing or giving feedback to each other.

C. Kim's (1988) study on the reading test practices in general English programs at 109 universities revealed that 35% of the teachers used non-standardized tests and 65% used standardized tests. He found that the test construction consisted of 26.9% translation, 23.8% reading comprehension, 20.2% grammar and sentence structure, 16.3% vocabulary, and 12.9% other. His study stated that most of the test construction practices did not meet the course objectives and had low reliability and validity.

K. Pyo's (2003) qualitative study on test practices in general English programs at universities revealed that the majority of the teachers interviewed used written tests, such as multiple-choice questions, and the rest of the respondents used other types of assessments. The study revealed that the reasons why the respondents preferred multiple-choice questions were because they were easy to mark and to analyze the results with the large number of students. K. Pyo's (2003) study has provided a deep understanding of classroom assessments through teacher interviews. However, he has not analyzed test papers specifically.

E. Kim (2005) investigated reading test construction practices in a general English program at a university to find out if the tests complied with the program objectives. She collected teachers' final tests and the syllabus, and analyzed the tests in terms of the number and type of test items. Her study showed that the tests varied widely and the majority of the teachers failed to meet the goals of the program. She stated the limitation of the study was that she was not able to identify the purpose of the tests and the problems that the teachers could have had in making tests due to insufficient data. She concluded that the study could have presented better results if teacher interviews were added.

More recently, K. Pyo (2008) conducted a survey to 86 English major students, through questionnaire and interviews, to understand their perceptions on English teachers' assessment practices in university. His findings indicated that students expressed concerns with teachers' low quality of tests, absence of communication channels for checking test results and grades with teachers, and undemocratic testing atmosphere. He stated that views and perspectives could vary depending on who the test stakeholders are and more attention should be paid in language testing.

Although the above studies have provided valuable insights into test practices in classrooms, none of the above studies have examined the test construction practices of NSETs in depth. With the large number of NSETs teaching at universities, it is necessary to examine their test construction practices to find out if they really enhance communicative competence (Canale & Swain, 1980). Thus, this study will examine native teachers' tests in English conversation classes to see if they assess students

effectively and comply with the program goals. Also, we examined their concerns regarding test construction for English conversation classes through a teacher survey.

III. METHOD

1. Participants

We used a case study approach to investigate the native teachers' test construction practices in English conversation classes. During the fall semester of 2016, the researcher selected seven native English speakers who were on the teaching staff at a university located in Gyeonggi province as the participants of this study (Table 1). Overall, the four male and three female participants ranged from 30 - 40 years in age. Their experience in teaching EFL varied from one to ten years, with an average time of instruction in Korea of 4.7 years. Participants represent a wide variety of nationalities, experience and educational backgrounds. They were from America, England, Canada, New Zealand, and Ireland. Only two have Master's degrees in TESOL, one has a Master's in Education and the others have Master's in various related fields.

TABLE 1

Demography of Participants

Name	Nationality	Gender	Major: Bachelor/Master	Teaching Experience
DC	USA	M	Economics/ Education	2 years
LL	USA	M	Psychology/ MBA	5 years
DS	England	M	Writing/Creative Writing	10 years
AH	Ireland	M	Sociology/ TESOL	4 years
DL	Canada	F	French/Education	9 years
EK	New Zealand	F	Asian Studies/TESOL	2 years
EC	USA	F	English Literature: Education	1 years

2. Overview of the General English Program

The English conversation course consists of two classes: English Conversation I and English Conversation II. The levels are intended to build upon one another. Students who complete English Conversation I are prepared to move on to the next level which is English Conversation II. The conversation classes are compulsory courses of instruction for all sophomore students at the university. The goal of English Conversation II for this

study is to enhance their communicative competence in English with a focus on speaking. Students enrolled in the class will learn about various topics and conversation skills using basic grammar and vocabulary. The course is 15 weeks long, and the class met once a week for 100 minutes. Class sizes ranged from 9 to 20 students per class. The curriculum is partially standardized, using the same textbook and syllabus but different mid-term and final tests. The course evaluation is as follows: 15% participation, 10% homework, 15% project, 30% midterm, and 30% final. The main textbook for English Conversation II is *Master Debate*. However, teachers are allowed to use supplementary materials in classrooms. Students enrolled in English conversation classes are assessed and evaluated by their teachers using non-standardized tests.

3. Course Book Description

The course book *Master Debate,* which is written and developed by one of the researchers, combines topics, themes, and talking tips. The book mainly focuses on speaking, however it also covers the skills of reading and writing in addition to vocabulary and idioms. The book has 13 units divided into topics, and each topic has a theme. Each unit has *Icebreakers* in which students can talk about the topic freely and brainstorm words and expressions related to the topic. The *Vocabulary Preview* and *Idiom Time* present new vocabulary and idiomatic expressions. The *Conversation* part has talking tips which help students speak like a native speaker. It uses a natural dialogue that introduces the vocabulary and idioms. In the *Topic Story* section, students can read interesting stories related to the unit's topic. In this section they can write and speak to share their own ideas and opinions about the story. The *Dynamic Discussion* gives students opportunities to discuss the questions related to each topic and provides a chance to debate topical issues.

4. Data Sources and Analysis

The purpose of case studies is to arrive at a comprehensive understanding of the groups under study and to develop general theoretical statements about regularities in social structure and process (Becker, 1968). This study is case research, where the data has been collected through test papers and a teacher survey (Appendix A) was conducted to support the test papers. The test papers were used to gather information about the teachers' test construction practices, such as the number of questions, and the types of questions. The survey had ten questions and was a written questionnaire, with both closed-ended and open-ended questions. In closed-ended questions, participants were asked to choose one answer among given choices. For open-ended questions they were

asked to express their opinions. The survey questions were designed to gather information about participants' thoughts and opinions on current test construction methods and to provide suggestions for development. The seven participants were asked to submit their mid-term and final test papers with the survey via email at the end of the fall semester of 2016.

5. Data Analysis

The data, including the mid-term and final tests of the seven native teachers, was analyzed in terms of the types of assessments and methods of the tests. The data for the surveys was not analyzed statistically due to a small sample size. Also, since both the mid-term and final tests for each teacher were in the same format and did not show significant differences, these tests were not analyzed separately. Referring to E. Kim's (2005) research, the data was categorized by the researchers into the following parts: types of assessments; number of question types; number of test items, types of language skills, and task types. The types of language skills were categorized into the following parts: vocabulary, grammar, composition, and reading comprehension. Then, the task types were categorized into the following parts: multiple-choice, cloze test, short-answer, opinion, and other. After analyzing the data, each teacher's test construction was compared and examined to see if they met the course goal, which was to enhance communicative competence. Also, the tests were analyzed to see if they were designed to meet the characteristics of the general program, which aims for standardization. The classroom project, which had students give a group-presentation accounting for 15% of their course grade, was not analyzed as it was done through a standardized process. All of the students who took English Conversation II had the same project and were graded based on the same rubric.

IV. RESULTS AND DISCUSSION

1. Analysis of NSETs' Test Construction Methods

The mid-term and final tests were gathered to discover the participants' test construction practices. For the purpose of research, the test questions were analyzed to investigate what and how teachers assess their students in classrooms. Among the seven participants, four teachers assessed the students through written tests (pen-and-paper test), two teachers used oral tests (performance test) and one teacher used both oral and written tests (Table 2).

TABLE 2

Assessment Types of Participants

Participants	DC	LL	DS	AH	DL	EK	EC
Oral					✓	✓	✓
Written	✓	✓	✓	✓			✓

The result of the analysis of the test papers showed that each teacher used different assessment types and each test varied in terms of numbers and question types. The number of question types used varied from 2-3 and the number of test items used per teacher varied from 20 to 60 (Table 3).

TABLE 3

Number of Question and Question Type Used in Written Test

Participants	DC	LL	DS	AH	DL	EK	EC
No. of question type	2	3	2	3	-	-	2
No. of test item	60	54	35	39	-	-	20

This shows that the teachers employed different numbers of questions on their tests, which could result in unfairness in scoring as there are different standards for score patterns across the tests. For example, teacher (EC) may have each question count for a larger portion of the test grade than teacher (DC) because she has 20 test items whereas teacher (DC) has 55 test items. E. Kim (2005) has reported that the number of test items used by teachers in her study ranged from 7-37. It is clear that our study also shows a large variation in the test item numbers.

Next, the test papers were analyzed based on the four language skills tested; vocabulary, grammar, composition, and reading-comprehension. Table 4 displays the frequency of language skills used.

TABLE 4

Types of Language Skills Used in Written Test

Participants	DC	LL	DS	AH	DL	EK	EC
Vocabulary	29	18	15	24	-	-	-
Grammar	0	0	0	0	-	-	-
Composition	31	17	0	7	-	-	-
Reading -Comprehension	0	19	20	8	-	-	1

Among the four skill types analyzed, vocabulary and composition questions were used the most on written tests. Reading comprehension was another skill that was used frequently, but grammar was not as frequently by native English teachers. However, the skills tested did reflect the skills in the course textbook, which did not contain a grammar section.

The test papers were also analyzed based on the four task types; Multiple-Choice, Short-Answer, Cloze Test, and Opinion. Table 5 presents the frequency of each type.

TABLE 5

Task Type

Participants	DC	LL	DS	AH	DL	EK	EC
Multiple-Choice	6	8	10	0	-	-	20
Short-Answer	17	45	0	0	-	-	-
Cloze Test (Gap filling, Matching)	23	0	15	24	-	-	-
Opinion	8	1	0	6	5	5	-
Others(True/False, Correcting Errors, Role-play)	6	0	10	9	1	1	1

The results showed that the teachers mostly utilized Short-Answer questions and Cloze Tests followed by Multiple-Choice items and Opinions. Others such as Sentence Creation, Picture Description, and True/False were also employed.

The above tables provide an analysis of the test methods, task types, and question types used by NSETs. The implications and the problems with the tests are analyzed as below:

- Each NSET has administered different assessment types. Four teachers used written tests, two of the seven used oral tests, and one used both written and oral tests. The aspects of test methods have great influence on language tests. Some test-takers assessed in speaking and some assessed in writing may cause inequity and unfairness because some test-takers may do well in speaking whereas some test-takers may do poorly in speaking. Thus, using non-standardized tests among teachers could result in unfairness as different test methods are employed depending on the teacher.
- Four out of seven test papers were pen-and-paper tests. Written tests are considered to be an inappropriate tool to assess conversation classes as they do not display learners' conversational skills in real situations. In the course syllabus, communicative competence was emphasized as the core skill to be enhanced,

however, assessing students through written tests might not aid in developing their real-world communication skills. So, it is implied that this type of assessment is not very reliable for assessing students' speaking skills.

- NSETs used questions where test takers could express their real opinions rather than grammar focused questions. For example, composition questions were commonly employed by the NSETs. These questions indicate that students have chances to freely express their own opinions about the topics. Still, language focused questions, such as vocabulary questions, were the most frequent question types on the majority of the tests. Of the teachers who employed written tests, most used vocabulary questions for almost half of their test items. Vocabulary questions, which are discrete points testing items, focus only on linguistic knowledge, but not on communicative competence.
- Multiple-choice questions, which are the most common type of discrete point testing, were the most common question types on the majority of the tests. Multiple-choice questions are used pervasively due to convenience and efficiency when administering and scoring. But, since they only require picking out one item from a set of given alternatives, they are not effective in testing the productive skills of speaking and writing as students do not create any language. Thus, these assessments are most appropriate for measuring receptive skills like listening and reading (Brown & Hudson, 1998).
- Another test item which was frequently used was the cloze test (gap-filling, matching). In the past, cloze tests were seen as an appropriate substitute for a test of productive skills because it required readers to integrate grammatical, lexical, contextual, and pragmatic knowledge in order to be able to supply the missing words. Also, a number of researchers voiced that it is a valuable tool to measure proficiency skills, if used correctly (Darnell, 1968; Stubbs & Tucker, 1974). But, further work showed that it is similar to a discrete point test of grammar and vocabulary, and does not test communicative skills (McNamara, 2000). Thus, using a cloze test also might not be suitable to measure students' communicative skills.
- There were inconsistencies in the questions. For example, Teacher A and B's tests were as follows.

Teacher A: *Q. The businessman earns a nice _____.*
a. open mind b. busy work c. chunk of change d. chores
Teacher B: *Q. I have too many books; they are sitting in a _____ pile on my desk.*

Teacher A used multiple-choice questions in which students were given four choices to choose the best answer. Unlike Teacher A, Teacher B did not present

any possible responses therefore it is more difficult for Teacher B's students to answer the question.

- There were inconsistencies in the score patterns across the tests. For example, Teacher C allotted 2 points, Teacher D 1 point, and Teacher F 5 points for the vocabulary matching questions.

Q. Draw a line from each word on the left to its correct meaning on the right.
Plethora *over-abundance, much more than we need*

Having inconsistencies in score patterns across the tests may cause unfairness to test-takers as the results of their tests could differ depending on which teacher they are assessed by.

- Rating is problematic because of different scoring criteria. Teachers who employed oral tests used their own grading rubrics. Each teacher had different marking criteria and did not use agreed upon criteria for judging the competencies. Thus, test takers experienced rater effect on their scores depending on who their teachers were. For example, DC used explicit components such as fluency, grammar, communication skills, and grammar, however, EK used descriptions that describe the test takers' performances in words. Because students were graded on different rating scales depending on the raters, it can be assumed that the ratings were subjective. Also, with one rater, there is no way to correct the mistakes or problems that could have arisen during tests.
- Of the teachers who implemented oral tests, they have shown a lack of reliability or validity on their tests. Reliability and Validity play important roles in performance tests (Brown, 1996; Heaton, 1990; Luoma, 2004, Underhill, 1987; Ur, 1996). There was only one rater and none of the tests were watched or observed by more than one rater.
- Washback (Alderson & Wall 1993) effect is another issue to consider. Washback is the effect of testing and assessment on the language teaching curriculum. With the current test methods, negative washback effect could be created. If the assessment procedures in a curriculum do not correspond to a curriculum's goals and objectives, the tests are likely to create a negative washback effect on those objectives and on the curriculum. For example, on the program syllabus, communicative performance was set as a goal of the course but most teachers assessed their students at the end of the courses with pen-and-paper tests. Canale and Swain (1980) demonstrate that pen-and-paper tests now in use do not necessarily give a valid indication of second language learners' skills in performing in actual communicative situations. Therefore, students may not work to cooperate

with the curriculum and its objectives. This could cause students to ignore the curriculum that is not directly related to the test.

2. Teacher's Opinions and Suggestions on the Current Test Construction Practices

The survey and the interviews for teachers provided feedback on the teachers' thoughts and opinions on their current test construction as well as a guide for future development of general English programs. This highlights the native teachers' perceptions of test construction and their ideal testing practices.

In response to the item "*What is the best way to assess the students in English conversation courses*?" two of the participants, AH and DC, answered that the best way to assess students in English conversation courses is by using written exams. They reasoned written tests are more reliable and can avoid rater bias. Although they preferred written exams, they admitted that it is necessary to assess students' speaking for specific reasons. AH said *"The best way to evaluate how much a student has retained material from the semester is through a written exam. However, as it is a conversation class it is important to evaluate the students' ability to communicate."* Five of the seven teachers, DL, DS, EC, LL, EK, asserted that the students should be assessed using both oral and written exams. One of the reasons was *"if a student is already a natural speaker, then they will almost always perform better than that of an unnatural speaker."*

In the question asking for their preference on the percentages of oral tests in conversation courses, AH, DC, EC, LL stated that 10%-30% is adequate, and DL, EK considered 0-50% adequate. Only DS responded that the percentage of oral exams should be more than 70%. The specific reason was "*...because speaking is the only way to assess students' true communicative abilities*.

The survey response to the item "*What do you think an oral exam should be like?"* DL, LL chose a 1:1 interview and one teacher chose a group presentation as an appropriate method to measure speaking skills. However, AH, DC, DS, EK stated that students should be assessed through various types of oral exams such as a 1:1 interview along with a solo or group presentation. DC gave a reason that "*...knowing presentation skills is beneficial because students will need to be able to make presentations later in life.*"

Thus, the results showed that the teachers believe oral exams should be employed in various ways. It is interesting to note that all the teachers agreed that students should be asked both textbook and non-textbook questions on oral exams. DL responded, "*Students should be tested by actual conversation skills in use.*" However, on the other hand, one teacher, AH raised a problem of employing test items from non-textbook by

saying "*...testing materials should be relevant to the materials covered in class as it allows the weaker students the opportunity to prepare for the test. If students are asked non-textbook questions, weaker students will not perform well*." It seems that oral exams should maintain both textbook and non-textbook related questions upon the students' agreement.

When asked the preferred frequency of tests per semester, AH, DS, EC, LL responded twice (mid-term and final test) a semester are sufficient, and EK, DL, DC reported three times (mid-term, final and a progress test) are adequate. It is notable that teachers do not show significant differences in frequency of tests per semester, yet, they suggested the percentage value on mid-term and finals should be lowered. Having a large portion of the grade on tests may cause students to only focus on their tests but not on their class participation. As DC pointed out, speaking should be assessed throughout the whole semesters. He said "*I think assessing students' speaking ability should be an ongoing process over a large sample size. It shouldn't be done in just one test day*." And also, teachers suggested to increase the participation score. By increasing the participation mark, students will actively take part in class activities to gain higher scores and to develop their speaking skills.

In the question asking for suggestions to improve the test construction methods in general English programs, teachers mentioned standardization of tests in format, regular staff meetings, and to form a Professor Committee to make mid-term and final tests as solutions. First, one of the seven teachers indicated that teachers should use standardized tests which everyone agreed upon. DL responded, "*All teachers should participate and give opinions on making test items and use the standardized tests and the marking rubric*." However, AH, DC, LL argued that teachers should have some freedom in making exam contents. AH reasoned that '*…it is an instructor who knows the best method for their class, therefore they should be able to apply it on their test construction*." Moreover, LL observed "*Teachers know their students' skills, abilities and interests, therefore they should make exam content individually. However, it is understandable to use the standard format.*" Another opinion that was generated by a DC was "*When I make test papers, I want to employ materials I teach in my class which is outside the scope of the textbook because I want my students to focus in my class.*"

V. CONCLUSION AND IMPLICATIONS

This study sheds light on the test construction practices of seven native teachers and also examines the results of a survey and the interviews designed to explore their opinions and rationales for their current test construction methods. To conclude,

suggestions for successful test construction methods are provided. These suggestions can be used by NSETs in university classrooms as an important reference when coordinating testing with the goals of the program.

The results indicate that despite having a standardized course goal and using the same textbook, the testing methods utilized by teachers varied widely. Firstly, most of the tests were in written form despite the course goal of enhancing communicative competence. Most of the students were assessed through discrete point testing, such as multiple-choice questions and cloze tests covering vocabulary. In addition, the tests were employed according to the teachers' discretion, with the number of questions and items tested varying from teacher to teacher. This can result in unfairness in scoring as there are different score patterns across the tests.

Furthermore, the teachers who employed speaking tests did not have validation of rating procedures in their assessments. The tests were scored by only one rater, and not watched or observed by a second rater, which can lead to rater bias. Also, using an unstandardized grading rubric can influence students' grades depending on their instructors. Lastly, negative washback effect can be created when the assessment procedures in a curriculum do not correspond with the curriculum's goals and objectives.

Based on the results of this study, and in light of current literature, we posit the following suggestions for the college English conversation courses:

First, teachers must base tests on clear definitions of both the abilities they wish to measure and the means by which they will assess these abilities. If the course syllabus suggests English class should be communicative, teachers should assess students using communicative methods. As supported in the literature, pen-and-paper tests are not a satisfactory way to measure students' real-world communicative skills. Thus, teachers should employ speaking tests in which students are provided sufficient opportunities to speak.

Second, standardizing exams in format, through regular staff meetings, and by forming a Professor Committee can be implemented. For example, universities can form a Professor Committee to make and develop a test format through regular staff meetings. In this way, not only a few members are involved, but everyone is taking part in standardizing the test and adding opinions. It is important to consider the test methods, items, number of questions, and marking rubrics to be used for grading.

Third, while it is true that speaking tests are an appropriate tool to measure students' communicative competence, these tests are admittedly difficult to administer. One proposed method to combat this difficulty is to swap a midterm exam for a speaking performance test. The speaking performance test is a gradual project completed throughout the semester in groups. Students create a presentation related to the course material with speaking checkpoints along the way. The culmination of this project is a

presentation in class. However, students will have to meet with professors one-on-one at least once throughout the semester to discuss their project. This enables professors to test a range of speaking items within one project. As one teacher suggested in the teacher interviews, this will give students more opportunities to develop their speaking skills gradually. Of course, there are other suitable methods to test speaking, but whatever is selected should include a few different assessments. For example, face-to-face interviews, individual presentations, group discussions, or role plays are all valid options.

It should be noted that this study has methodological limitations which should be addressed in future research. The weakness of this study lies in its data collection process. Grounded theory recommends that a researcher be constantly engaged in data collection and data analysis procedures to reach a saturation point. This research had only seven participants and thus with such a limited pool, the findings cannot be generalized and presented as fact. However, considering the scarcity of research on NSETs' test construction practice, we believe this study contributes to gaining a deeper understanding of NSETs' testing practices and their thoughts on testing through a qualitative lens.

REFERENCES

Alderson, J. C., & Wall, D. (1993). Does washback exist? *Applied Linguistics, 14*(2), 115-129.

Alderson, J. C., Clapham, C. M., & Wall. D. (1995). *Language test construction and evaluation.* Cambridge: Cambridge University Press.

Bachman, L. F. (1989). Assessment and evaluation. *Annual Review of Applied Linguistics, 10*, 210-226.

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

Bachman, L., & Palmer, A. (1996). *Language testing in practice.* Oxford: Oxford University Press.

Becker, H. S. (1968). Social observation and social case studies. *International Encyclopedia of the Social Sciences, 11*, 232-238.

Brown, J. D. (1996). *Testing in language programs.* Upper Saddle River, NJ: Prentice-Hall.

Brown, J. D., & Hudson, T. (1998). The alternatives in language assessment. *TESOL Quarterly, 32*(4), 653.

Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches of second language teaching and testing. *Applied Linguistics, 1*, 1-47.

Chang, Hyung-ji. (2009). A study on curriculum development for general English for the ABEEK program. *English Language & Literature Teaching, 15*(1), 327-348.

Chapelle, C. (1988). Field independence: a source of language test variance? *Language Testing, 5*(1), 62-82.

Chapelle, C., & Roberts, C. (1986). Ambiguity tolerance and field independence as predictors of proficiency in English as a second language. *Language Learning, 36,* 27-45.

Cho, Dong-Hwan. (1997). Study on item writing strategies for English tests. *English Teaching, 57*(2), 365-394.

Choi, Hoon, & Kang, Jung-Hye. (2009). Native English-speaking teachers, EFL education system, college student's English learning, English conversation class, EFL student. *The Association of English Language and Literature in Korea, 35*(1), 295-321.

Darnell, D. K. (1968). *The development of an English language proficiency test of foreign students using a clozentrophy procedure.* Boulder: Department of Speech and Drams, University of Colorado.

Fulcher, G. (2010). *Practical language testing.* London: Hodder Education.

Heaton, J. B. (1975). *Writing English language tests.* Harlow, England: Longman.

Heaton, J. B. (1990). *Writing English language tests*. London: Longman.

Henning, G. (1987). *A Guide to language testing: Development, evaluation, research.* London: Newbury House.

Hughes, A. (1990). *Testing for language teachers.* Cambridge: Cambridge University Press.

Jeon, J. (2010). Issues for English tests and assessments: A view from Korea. *Language Assessment in Asia: Local, Regional or Global, 4,* 55-82.

Jeong, Dong-soo, & Kim, Hae-Dong. (2001). A case study of general English course evaluation based on leaner's needs analysis. *English Teaching, 56*(4), 265-292.

Jeong, Yu-Jin. (2004). *The Influence of native speakers on university students' affectivity in English conversation classes.* Unpublished master's thesis. Jeonjoo University. Jeonjoo.

Joh, Jeongsoon. (2002). Towards a more successful college English program: A survey of teachers and students. *English Teaching, 57*(2), 365-394.

Kim, Choong-Bae. (1988). Proposal for testing freshman English reading. *English Teaching, 35*, 33-82.

Kim, Eun-Jeong. (2005). The teachers' test construction practices in a college English program: A case study. *English Teaching, 60*(2), 267-283.

Kwon, Oryang. (1995). A history of English teaching methods and methodology research in Korea. *English Teaching, 50*(2), 107-131.

Lee, Hwa-ja. (2001). The role of native English-speaking teachers in the Korean EFL education system. *The Study of English Education*, *6*(2), 33-67.

Lim, Byung-Bin. (1995). Problems in high school English teacher-made tests. *English Teaching, 50*(1), 1-20.

Lim, Jeong-Wan. (2006). A case study of general English course evaluation based on a questionnaire. *Modern English Education, 7*(2), 130-149.

Luoma, S. (2004). *Assessing speaking*. Cambridge: Cambridge University Press.

McNamara, T. F. (2000). *Language testing*. Oxford: Oxford University Press.

Park, Joon-Eon. (1997). Communicative freshman English teaching program by native speaker teachers. *English Teaching, 52* (1), 161-187.

Pyo, Kyung-Hyun. (2003). A case study of general English course evaluation in classrooms. *Foreign Languages Education, 10*(2), 273-290.

Pyo, Kyung-Hyun. (2008). Students' perceptions on English teachers' assessment practices in university. *Foreign Languages Education, 15*(1), 23-44.

Stubbs, J. B., & Tucker, G. R. (1974). The close test as a measure of English proficiency. *The Modern Language Journal*, *58*, 239-241.

Underhill, N. (1987). *Testing spoken language.* Cambridge: Cambridge University Press.

Ur. P. (1996). *A course in language teaching: practice and theory*. Cambridge: Cambridge University Press.

Yang, Taesun. (2011). Korean university students' perceptions about native and non - native English speaking teachers in TEE course. *English Language & Literature teaching, 17*(3). 237-254.

APPENDIX

Teacher Survey

1. In your opinion, what is the best way to assess the students in English Conversation courses?
   ① Only written ② Only oral ③ Both Written and oral
2. What do you think the percentage of oral exams in English Conversation course should be?
   ① 0% ② 10% ~ 30% ③ 30% ~ 50% ④ 50% ~ 70% ⑤ more than 70%
3. What do you think an oral exam should be like?
   ① 1:1 interview ② Solo presentation ③ Group presentation ④ All of the above
4. What kind of questions do you think students should be asked in an oral exam?
   ①Textbook questions ② Non-textbook questions
   ③ Both textbook and non-textbook questions

5. How often do you think you should assess your students?
   ① Only once (final exam) ② Twice (mid-term and final exams) ③ Three times (mid-term and final exams and a regular assessment) ④ More than four times
6. What types of assessments do you think are suitable for your written exams in conversation courses?
   ① Multiple-choice questions ② True-False questions ③ Short-answer questions ④ All of the above
7. What skills do you think are suitable for assessing your students in conversation courses?
   ① Listening ② Reading ③ Speaking ④ Writing ⑤ All of the above
8. Currently, how do you assess your students and why?
   ① Oral ② Written ③ Oral + Written
9. How can you make the oral test valid and reliable?
10. Please write down any suggestions you have for changing/improving assessment.

Examples in: English
Applicable Language: English
Applicable Level: Tertiary

Mee Hye Kang
Department of Liberal Arts
Eulji University
Sanseongdae-ro 553, Seongnam-si
Gyeonggi-do, 13135, Korea
E-mail: unipingu@gmail.com

Daniel James Cuffey
Department of Liberal Arts
Eulji University
Sanseongdae-ro 553, Seongnam-si
Gyeonggi-do, 13135, Korea
E-mail: daniel.cuffey@gmail.com

Myeong-Hee Seong
Faculty of Liberal Arts
Eulji University

Sanseongdae-ro 553, Seongnam-si
Gyeonggi-do, 13135, Korea
E-mail: seong@eulji.ac.kr

Received in July 15th, 2017
Reviewed in August 26th, 2017
Revised version received in September 5th, 2017