A Comparative Study of Korean University Students before and after a Criterion Referenced Test

A research project submitted by
Terry Joseph Frain,
B.A., M.A. (University of Southern Queensland, Australia)
In partial fulfillment of the requirements
For the award of
Master of Education
2009

ACKNOWLEDGEMENTS

I would like to deeply thank my supervisor, Dr. Shirley O'Neill for her support and advice in the preparation and writing of the survey items. I would also like to thank her for her advice on how to organize and write the project results.

I would like to thank the students at Dong Seo University, who made this research project possible.

Contents

1. Introduction

1.1 Statement of the Research Question

A comparative study of university students and their perceptions of English language testing in Korea before and after a criterion referenced test. How do they perceive English language testing?

1.2 Statement of the Research Problem

Multiple choice exams are the main English language testing method in the Korean public education system. The researcher believes that they are limited in scope as they exclude assessment of speaking and writing abilities while only testing for listening and reading skills. Although the curriculum aspires to use Communicative Language Teaching (CLT) methodology, this may not be the case as the teachers are employing a more structural methodology that teaches to the test in reading and listening skills. Negative backwash seems to ensue as students are studying to become test wise while, at the same time, not developing communication skills in English. This results in poor scores for Korean students on tests of communication proficiency as students frequently try to memorize answers to multiple choice answers (Kang, 2008). Negative backwash from such testing methods may force some families to separate and go abroad to get an English language education where methodology and assessment are complementary. Poor scores from such multiple-choice exams that divide language assessment into separate skill focused sections may even result in suicide, such is the competitive nature of testing in Korea.

1.3 Context and Background of the Research Problem

The researcher was puzzled by the assessment method being employed and the negative backwash it was creating. He had studied a foreign language at university and was bewildered by the fact that the majority of testing does not seem to provide feedback to the student. An endless series of testing seems to characterize the Korean educational landscape and whose purpose remains dubious or wanting in an educational context. It appears that the selection of students for university entrance is the main purpose for studying English among young Koreans. The usefulness of the university entrance test is compromised as the test method segregates students by measuring their accuracy in language use, not their fluency. As a result, the input macro skills of listening and reading are preferred to the output skills of speaking and writing. By testing only half of the macro skills, the limitations of the testing method circumscribe the way the results can be interpreted, especially the output skills of speaking and writing.

This approach to language testing has other limitations. Similar tests such as the TOEIC have a bias toward testing for vocabulary and grammar. The development of proficiency skills is of secondary importance to the mastery of individual elements of language usage such as grammatical accuracy or vocabulary. Individual macro skills are divided into distinct entities to be tested separately. The TOEIC fulfills the role of classifying or ranking the students as it is norm referenced and provides a criteria in the job selection process. But it provides limited feedback for the student and, oftentimes, contribute to negative backwash. Characteristic of this type of psychometric testing is the multiple choice exam, which is regarded as objective in nature as it is free from the influence of subjective judgment.

This method of testing is in contrast to the methodology of language teaching that is currently being employed in Korea. Communicative Language Teaching (CLT), as the name implies, stresses fluency as opposed to accuracy in language use. With this goal, accuracy is judged in context and not as abstract entity. Fluency would imply that a variety of different answers are possible when two people communicate and that the evaluation process would have to reflect this inherent quality of language. This seems to imply that proficiency language testing would have to take into consideration how successful the candidate was in communicating a response to a task rather than language accuracy. Descriptive bands would have to used to measure the performance level of the candidate. They would provide a qualitative assessment of the student as opposed to a quantitative one. However, such a testing method would also imply that a subjective judgment, rather than an objective one, would have to be employed for assessment purposes.

A pilot survey was developed based on the qualities that make a language test useful, namely, reliability, validity, authenticity, interactiveness, practicality and impact. This survey was first translated into Korean and used on first year university students during their first semester. They had studied English from the third grade but many of them had limited experience speaking and writing in English because of the emphasis on listening and reading. Once in university, a shift in testing methods occurs from a quantitative to a qualitative assessment. This qualitative approach has distinct advantages. It evaluates students on their ability to perform a task rather than in relation to other students (normative testing). It also describes the student's strengths and weaknesses using the target language and provides feedback for the student so that they can benchmark themselves and determine their language ability. The researcher was interested in their perception of this new testing method, their attitude(s) toward it and whether this new method was providing positive backwash so they could score higher in tests of communication competence such as the IELTS, something they do very poorly in at the present time (Kang, 2008). This project compares the English language testing attitudes and perceptions of Korean students before and after a criterion referenced test to determine how they view language testing, its constraints and attributes and how testing could result in positive backwash so that the current methodology (CLT) and testing practices are more compatible.

1.4 Abstract

The study aims to determine the perceptions of first year university students to criterion referenced testing. The students have been tested using norm referenced testing for most of their English language education and this has culminated in The College Scholastic Aptitude Test (CSAT).The poor communication skills of the students has prompted the researcher to question why CLT methodology is not complemented by a communicative

test that reflects real life situations practiced in the classroom. The attitudes and perceptions of the students may support a different method of testing that complements a communicative approach to learning. It seems that backwash from the CSAT, which emphasizes only reading and listening, is negatively affecting communicative competence (Flattery, 2007). The experimental approach will be action research as this is a single case study of students and their attitudes and perceptions about language testing. It seeks to understand the effect that backwash from a test has on them. The students were tested using a paired criterion referenced test during their first semester at university. They are surveyed twice, before and after the criterion referenced test, to determine their opinions about this new testing method and norm referenced testing. The survey items reflected the qualities of a good language test, namely, interactiveness, practicality, reliability, validity, practicality, and impact. The results seem to indicate that the students question the reliability of norm referenced testing while criterion referenced testing created positive backwash. The students perceive the use of real world tasks as being more relevant in assessing their abilities in English compared to decontextualized multiple choice exams. They also perceive that they are no longer being compared with each other but in their ability to perform a task, which seemed to create a positive attitude toward language learning.

1.5 Definitions: Tables of Contents of Special Terms

Backwash. The effect that a test can have on a students, teachers, and society as a whole. It can be both positive and negative in nature.

Communicative Language Teaching (CLT). A contemporary language teaching methodology.

Criterion Referenced Test (CR test). A language test that measures a candidate’s ability to perform a task in a specific domain.

CSAT. The College Scholastic Aptitude Test. The one shot university entrance exam that Korean students must take to gain admission to university. It has an approximate 20% weighting in English but currently tests only listening and reading skills.

Direct Testing. A test used to measure the productive skills of speaking or writing that requires the student to perform a particular task.

ETS: a nonprofit institution advancing quality education with valid educational testing, curriculum development assets and test prep products.

iBT TOEFL: Internet based Test of English as a Foreign Language

IELTS: The International English Language Testing System

KICE. Korean Institute for Curriculum and Evaluation

Norm Referenced Test (NR test). A language test that relates one candidate's performance to that of other candidates.

Objective Test. A test in which the assessment is made without bias. There is only one correct answer.

SL. Second language

Standardized Test. A language test that that is administered and assessed in a consistent manner.

Subjective Test. A test in which assessment is arrived using an opinion. This type of test implies that there may be more than one way to answer a question.

TOEIC: The Test of English for International Communication.

2. Literature Review

2.1 Overview of Korean educational and assessment practices and their impact

Foreign language teaching may vary widely from country to country. In the case of Korea, English is a compulsory foreign language subject from grade three of elementary school until the end of the second year at university. The sixth and seventh language curriculums for elementary, middle and high schools of the Ministry of Education have advocated Communicative Language Teaching (CLT) as its preferred methodology for teaching English (Park, 2006). However, this methodology has not employed an assessment that recognizes fluency as opposed to accuracy in language use.

The norm in testing remains multiple choice exams that culminate in the college scholastic ability test (CSAT) for high school students and THE TOEIC for university students when they apply for a job. These multiple choice exams restrict what can be tested, are subject to guessing, and only recognize language forms at the expense of language function. This testing method leads to backwash or the effect that tests have on language teachers and learners to do things they would not otherwise do that promote or inhibit language learning (Messick,1996). This effect can be harmful to the learners, the teachers, the education system and society at large. The learners realize that they do not have to speak or write in English to get a good grade. They can simply concentrate on a correct answer by focusing on lower order knowledge or forms while the teachers are pressured to teach to the test. In this literature review the qualities of language test, namely reliability, validity, impact, practicality, authenticity and interactiveness will be discussed for the purpose of better understanding Korean assessment practices.

The educational authorities seem to support a multiple choice testing method as it is deemed objective and reliable while being practical at the same time. However, this method of language testing presents a paradox to the very authorities who introduced a communicative approach to language learning as fluency is not being measured. Because the purpose of testing is selection and not evaluation of student performance, feedback is limited to the resulting grades and the teaching practices that were used to obtain them. Neither the student nor society seems to question these one shot exams that determine the future of the student as this testing purpose (selection) seems to complement the nature of rank in a Confucian society. Confucian meritocracy is regarded as being responsible for the creation of these one shot exams that can determine a student's future. While having the goal of creating a classless society, meritocracy has instead created a flourishing duel education system. It is against this background that the teachers find themselves. As a result, a student is subjected to a very stressful environment that contributes to the high suicide rate among its youth (Breen, 2004).

While the teachers are compromised into teaching to a test and not a curriculum with its avowed methodology, the effect on students is the opposite. While the CSAT is of paramount importance, other exams in English loom on the horizon. The TOEIC test is used as a criterion for selecting job candidates and a communicative test of language ability will follow this if the candidate is selected in the application process. Thus, the student is approaching language learning for different purposes and probably employing different strategies. On the one hand, the student may be memorizing discrete grammar points at a cram school while studying fluency at another academy.

The emergence of standardized proficiency tests such as IELTS and iBT TOEFL would seem to be gaining more popularity in Asia, especially with the expansion of international programs in English at universities. In order for Asian countries to compete globally, governments are promoting these international programs at university as English proficiency, analytical ability, critical thinking and computer literacy will help them compete economically (Prapphal, 2008). Indeed, the Ministry of Education in Korea is considering introducing a national English proficiency exam based on Japan's Eiken test (Kang, 2008). While the primary purpose of this proficiency test is to create a single national test for all students, its implementation may have the opposite effect if it is not unanimously adopted.

These different tests, with their distinct formats may only create new and unforeseen sources of backwash for students, teachers and society. As an example, the Ministry of Education has for the first time released the regional results of the CSAT. It showed a wide performance gap among districts in Korea. These results could be used as a means to label students by the school or province in the country they come from. Private cram schools may use the results as a marketing tool to correct the imbalance that exists between good and bad schools. Parents may want to place students in better schools while others will become disconsolate, realizing that their local school performed poorly and that the chances of their offspring getting a proper education are dwindling (Bae, 2009).

2.2 Korean cultural considerations and their impact.

Korean society spends an average $11 billion on private English education each year (Kang, 2009). It may be fair to say that a considerable amount of a family's money is being spent on the CSAT as this exam has such importance and status that it can determine the future livelihood of many young Koreans. This payment for private tuition fees gives a distinct advantage to some students over others. Depending on currency rates it may be cheaper for a family to split up with the mother and child leaving Korea to study in Australia or Canada with the father remaining in Korea to work. The social cost of family separation that results is not directly tangible but it can lead to loneliness and a sense of inadequacy for the entire family. The fixation on university admission is probably related to the Korean penchant for ranking as students are entering university not only to acquire an education but to obtain social status. In a Confucian society, by obtaining status, you are creating order and promoting a harmonious society. This cultural trait seems to be so strong that Koreans will pay or do whatever it takes in order to obtain it (Breen, 2008).

Kim (2004) emphasizes the cultural barriers that a collectivist society imposes and possible solutions. In a collectivist society, uniqueness and initiative may be viewed as a vice and not something positive that can be used as a building block for language learning. Hence, a quiet student is viewed as one who is more respectful than one who engages in debate and critical thinking. The hierarchical nature of society also reinforces this precept as age and gender determine who can express opinions and dispense knowledge. This cultural practice permeates peer assessment also as students may resist offering a correct answer after a mistake is made out of mutual courtesy. Error correction, which can be viewed as positive method in language learning, is not commonly practiced in a Korean classroom. Thus, a student is likely to only speak up in class when invited to do so by the teacher. While Kim's views on the cultural dilemma that English teachers face is relevant, he offers very few suggestions to correct passive learning. A few interactive activities in a Korean context seem to be his only solution.

This article is relevant in the context in which Korean students finds himself. The hierarchical nature of society is embedded in a student's mind at an early age. A young child enters into an education system that constantly tests. By constant testing, I am referring to weekly or unit by unit chapter testing. An inordinate amount of class time that could be used to develop language skills is seemingly lost to this practice. Frequent tests may be used not only to diagnose student problems but as a means to determine if the teacher's method was effective. But, in a Korean context, these classroom tests seem to serve as a mechanism to rank someone and not to provide feedback. A young student learns what their 'position' is and, as a consequence, their motivation and desire may suffer. This may be because the poor proficiency levels of the teachers restrict them in their ability to provide relevant feedback (Kim, 2003). Because formative assessment is not supplying feedback, it does not distinguish itself from summative assessment, which provides a student with a grade at the end of the semester. Both types of assessment should be employed in the classroom to provide a more comprehensive evaluation of the student (Caridad, 2009). In any case, students seem to be always engaged in exams which seem to be of limited value and a waste of valuable class time.

In Korea, critical thinking seems to take a back seat to rote memorization and regurgitation as the education value system is more interested in arriving at a correct answer than determining a better approach to epistemology (Stevens, 2009). An alternative sociolinguistic approach to learning a language which may be appropriate is intercultural language learning (ILL) proposed by Liddicoat. (Liddicoat,1999). This approach creates a third culture that examines the constructions of one's own culture and the language target culture so that cultural awareness is activated and students are able to become comfortable participants in a third place or culture. The characteristics of this third culture would foster a learning strategy whereby a learner creates meaning, questions social categorizations, encourages making connections to dominant attitudes, and is contextually sensitive. As a result, language learners have a better understanding of the target language and culture (L2 and C2). This intercultural competence contributes to successful communication across language barriers (Kramsch, 2008). Practical applications in a Korean classroom would be difficult to envisage. It would require that the students be empowered to explore and investigate new ways of looking at their own culture and others. Further, due to the restrictive nature of objective tests, students would not have a chance to perform a given task or solve a problem using their linguistic skills.

Instead of learning how to develop in a co-operative environment, Korean students may go from being friends with each other to competitors who are endlessly sitting a series of exams geared towards getting the correct answer. The strain results in one of the highest suicide rates in the world, something that rarely gets mentioned in the Korean press. Suicide prevention programs, which are used in some countries as a mechanism to prevent suicides, are not popular in Korea (Ruffin, 2009). At the very least preventative education should be implemented and follow–up counseling provided so that the grieving families of the victim don't feel guilty for the rest of their lives.

Flattery (2007) notes that the majority of high school and university students believe that English is necessary to be successful in an age of globalization but that the materials employed fall short of obtaining this goal. Many university students must pass company interviews in English to get a job. But they complain there is a lack of student interaction and student-oriented activities, which would help improve their communication skills. There is an emphasis on reading and listening in the classroom at the expense of speaking and writing. He also notes that both teachers and students seem reluctant to use a student centered approach to learning. He does not elaborate on this point however. The CSAT does not include speaking and writing so this seems to create passive learners who are not interested in student centered activities. Teachers would be emphasizing listening and reading macroskills and concentrating on improving test taking skills such as absorbing vocabulary and grammatical detail. Perhaps there is still a belief that that by testing grammar and vocabulary, proficiency will naturally follow, but this belief has been proven erroneous (Brown & Hudson, 2002). It is more probable that these skills are tested because they are easier to teach and grade, given the proficiency level of some teachers. The other issue Flattery doesn't deal with is the negative effect these one shot tests have. There is a tendency for teachers to teach to the 'average' group in a classroom to the detriment of those more or less gifted (Henning, 1988). The teachers have little or no feedback outside of the test scores and are thus limited in determining how to better prepare students for these tests.

It should be recognized that the negative backwash created by the CSAT is not particular to Korea alone. In Thailand, for example, a multiple choice English language test is part of a larger exam used to determine university entrance due to its practicality and reliability. Cram schools remain the norm. However, the Ministry of Education in Thailand has recently moved to a quota system to determine university admission. It has also set up a special admissions project to help the rural poor (Prapphal, 2008).

Classroom based assessment is a possible solution to this problem if feedback were provided as it enables teachers to make adjustments in their teaching to improve their own effectiveness so that more students can benefit from the learning experience. Classroom tests also enable the teacher to pinpoint areas of weakness that the students have and concentrate on them. Without these tests, the teacher may not be able to determine which parts of the syllabus are effective or whether the proper materials are being used. They also enable the integration of assessment and instruction, something that is occurring very slowly in Korea due to lack of training (Kim, 2003). For the student, it allows them to be tested in a less hostile environment and makes them aware that more than one correct answer is a possibility. They learn how to collaborate with one another, something that they will have to do later in life when they get a job. If their classroom assessment provided them with a descriptive scale, they would become aware of their strengths and weaknesses and, more importantly, of the progress they are making. A bi-directional or self-reflective approach to assessment would become a distinct possibility. This formative approach to assessment would be difficult in Korea as a huge shift from passive to active learning would have to occur. Compounding this would be the teacher's obligation to teach to the test, particularly at the high school level.

Flattery, (2007) mentions the cultural problems Koreans encounter learning a foreign language. He maintains that CLT must be culturally appropriate and that teaching practices must be changed to reflect this. While he does mention the limited nature of testing in Korea, he does not discuss the nature of achievement tests. It seems that achievements tests in Korea emphasize ESL content at the expense of EFL content (Kim, 2003). If this is occurring, then Korean cultural considerations are not being recognized in testing, which would seem to create negative backwash. There seems to be a movement for better teaching practices nowadays as cultural considerations are being written into the syllabus. If this is so, there seems to be no reason why achievement tests can't become more focused on creating positive backwash and less focused on discrete point accuracy using objective tests.

2.3 Assessment problems and practical solutions.

Finch and Shin (2005) compare many of the assessment problems and possible solutions in a Korean context. While one of the most popular features of a multiple choice exam is its practicality because of large class sizes, a possible alternative is the introduction of pair or group criterion referenced testing. This form of classroom based assessment (CBA) would have distinct advantages as it permits higher order learning outcomes and thinking skills to be employed compared to a multiple choice exam based on knowledge recognition. It may make the students more aware not only of the learning process but also of assessment and reduce the workload of the teacher by engaging the students in the evaluation process. Finch and Shin also mention that the topic of language assessment is not popular and few opportunities exist for professional development in this area. This seems to imply that if the educational authorities are going to introduce classroom based assessment, the teachers may not have the expertise to evaluate it. Compounding this problem is the poor language speaking and writing skills of the teachers (Kim, 2003). Finch and Shim do not mention the amount of time and money that would be required to improve the skills of the teachers to facilitate and administer such a testing method.

Finch and Shin correctly view testing and instruction as being separate in Korea whereas assessment should be an integral part of instruction. Korean tests can be regarded as norm referenced as they are developed independent of instruction and tend to limit the usage of authentic language. These standardized tests usually have a multiple choice format and are designed and developed to maximize distinctions between individuals. As a consequence, the students are motivated extrinsically to get the correct answer, and not intrinsically for the sake of performing a task and learning how to communicate in a different language. Motivation is an important tool in the language learning process as it can be independent of ability or aptitude (Gardner, 1985). A teacher must introduce language tasks that are stimulating and nurture intrinsic motivation. This teaching technique is characteristic of authentic pedagogy which complements the intellectual demands of the students. Authentic pedagogy empowers the student to look at values beyond school while minimizing the problems of traditional curriculums. This approach would imply that language learning and testing are constantly evolving, something that doesn't occur in standardized testing, which is linear, predictable and measurable (Finch and Shin, 2005). Thus, the current testing methods would have to change substantially to recognize the student centered nature of learning and assessment so that student demands of authentic pedagogy could be met. Intrinsic motivation could be further nurtured if the teacher provided positive feedback that is informative so students can assess their progress and problems on a more personal level.

If the students are to receive grades within an educational context, then any evaluation would have to be accountable (Bachmann, 1990). The issue of accountability is neglected by Shin and Finch as students may be advanced without consideration of their scores. This parallels language testing at the elementary level in Korea as it is informal, consisting of practice tests from TOEIC books for children (Ahn, 2003). There really is not a reason to test in such a situation as the results would not be meaningful and useful. In such a situation, students may resort to self-assessment to determine what their ability is. In a learner centered approach to education, they are encouraged to not only be test takers but also be active participants in the assessment process. While a learner's judgment is subject to variables such as lack of training and the amount of exposure to foreign languages, self-assessment may provide an approximate benchmark for the student when accountability is overlooked by educational authorities (Saito, 2009). This self-assessment could be on-line in the form of quizzes that closely resemble the current level and skills that students want judged. Besides self-assessment, the internet provides the opportunity for collaborative video conferencing. This new technology allows the students the opportunity to interact with peers from different cultures, share resources and compare their language skills with each other (Atkinson and Davies, 2000). This resource seems to be under utilized in Korea probably due to the novelty of collaborative learning.

Finch and Shin mention that active as opposed to passive learning would occur if the testing methods were to change. This is a valid point and would have a positive effect on intrinsic motivation. Teachers could use the strengths of the students as a motivational tool to fulfill classroom tasks. The students would soon realize that they are no longer competing but cooperating with one another as learners and would be assessed according to their own performances. The problem is that Korean classrooms are teacher centered and not student centered (Ahn, 2003). To state that learning styles would change once a teacher introduces a different type of test may be presumptuous as their role would likewise have to change from dominating to facilitating a class. They may feel threatened both professionally if they have poor language skills and culturally as their traditional role in the classroom would be changing.

Finch and Shin also advocate same grades for students in group tests. This method of testing has the advantage of having the students interact with each other and thus satisfy one of the possible demands of the test specifications. It may also elicit a better response by candidates as the examiner is not directly involved. However, if one candidate dominates the group, making generalizations about other members can become problematic (Hughes, 1989). This approach also seems to ignore the various personal factors that affect test scores such as cognitive styles, content area knowledge, superior ability in grammar, discourse and pragmatics. The validity of a test score could be suspect if everyone received the same grade. Uniform results may imply that students were tested on material that was easy to test, not on constructs that were easy to test.

2.4 Reliability in a Korean context

Reliability can be viewed as the similarity in test scores if the same student took the same test on different days (Hughes, 1989). Sewell (2005) delves more deeply into the quality of test reliability in a Korean context. He correctly analyzes the test–retest data for The TOEIC, a multiple choice exam, and deduces that a student's score can go up because of test wiseness, whereby a student develops test taking strategies that improve a test score in listening and reading. However, the problem that arises is that these improved scores do not co-relate well with the output macroskills scores of speaking and writing, which essentially remain the same. Thus, reliability in a Korean context would seem to be concerned with only the macroskills of reading and listening.

Test wiseness is of particular interest in Korea where many language exams have the same format. Students may try to rote learn answers or develop test strategies such as test wiseness to achieve a higher score on a multiple-choice exam. To take advantage of this situation ETS has produced a large amount of TOEIC preparation material, some of which is taught in schools and private institutes. This may infringe on the curriculum that attempts to use authentic language in a communicative approach to learning. It will be interesting to see how well the new TOEIC test, which includes speaking and writing, correlates with other proficiency exams. If it fails to engage the candidate in a series of different topics to determine if the candidate can speak and write at length about them, then its claim to be a reliable proficiency test might prove to be dubious (Knapman, 2007).

While multiple-choice exams are by nature objective, their repetitive use can have harmful effects. Besides guessing, they test only recognition knowledge and not functionality in language use. They also cause negative backwash by having students develop test strategies as opposed to developing functional language skills. With these negative attributes, it remains a mystery why multiple choice exams are so popular in Korean language testing. It may be because they measures only language recognition and not language use. If a proficiency exam measuring all four macro skills were adopted, it would seriously challenge the English test preparation industry in Korea. It is biased toward multiple choice exams and may not have the ability to change into an institute that teaches integrated macroskills for communicative language use for lack of qualified teachers. Rater reliability may also be an issue as the majority of teachers lack both training and proper assessment instruments (Kim, 2003).

If a Korean CSAT or TOEIC teacher were to be replaced by a foreign teacher or a second generation Korean American, this may cause a serious rift among the teachers, whether they be in private or public education. Taking someone's job off of them after years of dedication and commitment because they only specialized in certain parts of a language, would be a tough pill to swallow. Additionally, the language institute owner may have to recruit a completely different roster of teachers while explaining to anxious parents that he is doing this because of a change in language testing methods. His credibility as an institute director may be questioned and this could result in a loss of business. Indeed, certain groups may want to perpetuate the present testing method claiming it is objective and reliable. Such groups have used the present system and may feel that their superiority (status) is being compromised by a change. Their success may have been directly related to Korea's thriving private education system, something the present government is trying to eradicate. As a result, families have to compete with one another to provide their children with the best possible preparation for the CSAT.

Tuition fees for students to attend cram schools and take exams are expensive. Socio economic status may become an issue for students as the quantity of time spent for test preparation and the quality of test preparation can directly affect scores as strategies such as test wiseness would seem to indicate. Thus a student's success on a test may correlated to how well endowed his family is. While The Ministry of Education has allowed a limited amount of recruitment for university entrance to be done outside the CSAT, this has unleashed a backlash by private universities that want to stage their own admission exams (Kang, 2008). Still other groups advocate that the current weighting of the CSAT be lowered and that interviews and recommendations from principals receive more importance. While the Ministry of Education and private universities cannot decide on a college admission system for less privileged students, still others propose a more radical approach in which a more diversified selection system is used. This may include the student's entire portfolio and not just grades at high school. This approach stems from the belief that Korea already has enough students who can solve test questions but is missing out on other students who may be more creative (Choi, 2009).

While these factors may influence reliability, it is the interpretation of reliability in a Korean context that is important. This interpretation may have more to do with consistency at the expense of a test measuring what it is supposed to measure. Simply put, a language test becomes a device to segregate students while at the same time maintaining its consistency. If the test succeeds at this, then it is reliable and useful even though the content validity of the test may have been violated. It seems that Koreans are satisfied with this interpretation of reliability due to their penchant for segregating students in a consistent manner.

2.5 Validity in a Korean context.

If a test measures accurately what it intends to measure, it is considered to be valid (Hughes,2003).This general definition of validity has been refined into distinct parts. Construct validity relates to whether the test measures individual skills, abilities or attributes the student has acquired during the semester. This measurement, and the inferences to be made from it, are limited as students are only tested on listening and reading macroskills. Simply put, any abilities that a student has in speaking or writing are superfluous as the current test method does not test them. Construct validity may be further compromised in such a high stakes tests such as the CSAT as the test may require vocabulary and familiarity with grammatical structures the student was never exposed to (Henning, 1987).

Concurrent validity is determined by comparing results from one test format with those from another test which essentially tested the same instrument (Nall, 2003).If the test results are strongly correlated, then the results are considered valid. The problem is that test scores in the input skills of listening and reading don't correlate well with the output skills of speaking and writing. A student may be able to read a text and recognize the present continuous but not be able to use it in a speaking context. When this occurs, multiple choice test results become extremely limited in how and what inferences can be made about them.

Validity also is concerned with the impact a test has on society, individuals and the educational authorities (Bachman,1990). Korean society holds the CSAT in high esteem, not only for its importance as a gateway to university but also because of its reliability in determining which students are successful or not. This gives the CSAT face validity as it is generally recognized by the public as being valid. Students have prepared years for this one shot test. This preparation time has also consumed a fair amount of the family's finances. In a theoretical sense, the constructs tested should be characteristic of the curriculum that was used. While some may argue that this is the case, the pedagogical reality of test preparation dominates the latter years of a student's high school English education. These types of high stakes tests seem content to employ more decontextualized material that focus on grammar and vocabulary usage that was used during the last latter years of high school education. This helps to segregate students for selection purposes as the CSAT is heavily weighted towards language accuracy. On the other hand, a proficiency test would have explicit specifications about the constructs that are to be used. These specifications should test the student directly on the constructs and the scoring should reflect a student's ability on these constructs (Hughes, 1989). If this were the case, The Ministry of Education would have to adopt a proficiency model for examination purposes, so that the constructs being used in the classrooms were tested on the exam. This approach to defining language proficiency could take two forms, a real-life approach or an ability approach (Bachmann, 1990). The former would probably prove better in a Korean context as the domain of language use can be defined, something that is missing in decontextualized testing. Also, the teachers would be better able to assess real world topics they are familiar with.

Language teachers would have to learn the format of the new testing method to ensure objectivity. They would have to become familiar with band descriptors and how to interpret them. This would require that they practice so that their reliability can be ensured. They would have to show how to subjectively arrive at a mark using them. Rater reliability becomes of paramount importance as the candidates are no longer being tested against each other but against their ability to use a language. By using explicit criteria in the marking scales, recordings and multiple examiners, grading becomes more objective. In writing exams, by using one topic, the examiner can limit the way the candidate can respond and thus ensure more reliable grading for all the candidates. By using a blind marking procedure and knowing only a candidate's number, objectivity is promoted. Finally, by allowing an appeal process, objectivity is promoted in a proficiency exam. Korean teachers would need to grow in confidence and ability to adapt to a completely different testing method. They would also have to assume a more active approach to testing as opposed to their current passive role that is dependent on Ministry of Education standardized testing (Kang, 2009).

In a Korean context the question of subjectivity becomes clouded by one's definition of competence. Nunn (2005) believes that competence is an abstract term that is defined in relation to the communities in which it is implied. That would seem to mean that norm referenced testing, which is commonly practiced in Korea for selection purposes, may be more highly regarded for measuring competence than in another country. This has to do with face validity or the extent to which a test meets the expectations of those involved in it, namely the candidates, administrators, teachers and society in general (McNamara, 2000). It may also have to do with the fact that an acceptable level of achievement can only be determined after the test has been administered and through reference to the mean score of the student population (Mangubhai, 2006). This may tend to segregate students better than a proficiency exam in a Korean context because of their inexperience with proficiency testing. However, even in proficiency testing, through the use of descriptive bands and the employment of competent and experienced examiners, a precise assessment of language ability is available. What Nunn doesn't elaborate on is the use of international tests such as the

IELTS that try to standardize testing to measure proficiency. Kang’s (2008) comments that Koreans rank very poorly on IELTS tests would seem to indicate that their current language education and testing methods have been a failure. It may further imply that Koreans are aware of but unable to devise, construct and implement effective proficiency tests such is their fixation with objectivity and norm referenced testing.

2.6

Interactiveness and Authenticity.

Bachman and Palmer (1996) define interactiveness as the extent and type of involvement of an examinee’s individual characteristics in accomplishing a test task. These characteristics may include language knowledge, strategic competence or the ability to paraphrase or use circumlocution and topical knowledge. These abilities require that the candidate use critical thinking and creativity at times. Because a multiple-choice test has a fixed response format in which a candidate chooses one of a number of responses, it tests only recognition knowledge and does not necessarily reflect a model of language in use (McNamara, 2000). As a consequence, the candidates are required to get a correct answer and not on apply the rules and structures in real life situation as they are afforded no opportunity to show what they can or cannot do with the target language.

Authenticity is the degree of correspondence of the language test task to the features of target language usage (TLU) domain so that inferences can be properly determined (Bachmann and Palmer,1996).A language test task thus becomes a function of two qualities as a test taker must not only process schemata or topical knowledge but also use language knowledge. By using conventional testing methods, such as multiple choice exams, only lower order language knowledge is being tested in a limited way. This can be expected to change as electronic assessment becomes more popular using exams such as the iBT TOEFL. The issue of electronic assessment is a new field, not without its problems. A test taker can role play a hypothetical situation as a sales clerk with a potential customer (a partner or examiner) in a test that duplicates real life situations with actual participants. This assessment format is both authentic and interactive. It would be more difficult to envision using an electronic test, where there is no interlocutor present. The response of the candidate may suffer in key areas of language ability such as strategic competence and metacognitive strategies without the presence of another person. Besides interactiveness, authenticity may be limited. The perceived relevance of the test task may be questioned by the candidate if they are simply speaking into a microphone. This may not produce the expected response to the test task. This could affect the content validity of the test as the interpretation of scores in non test language use domains may be suspect. It would seem that self-assessment is a better area to use electronic assessment. However, this also has its limitations as the output skills of writing and speaking are generally considered more difficult to assess than the input macroskills of reading and listening.

Students generally score very poorly on tests of communicative proficiency as the testing method would also have to complement CLT, which has fluency, not accuracy, as its primary goal (Kang, 2008). In a Korean context, teacher judgment in a proficiency test should be adequate if it is made against valid benchmarks. This qualitative approach has two advantages. It introduces a more pragmatic approach to language testing that evaluates students on their ability to perform a task rather than in relation to other students (normative testing). It also describes the student's strengths and weaknesses using the target language. This approach may also help students score higher in tests of communication competence such as the IELTS.

2.7 Limitations of the proposal.

This research proposal has many limitations. The teachers must teach to a test and not a curriculum as the students prepare for the CSAT. This will probably continue for some time as KICE (Korean Institute for Curriculum and Evaluation) has no plans to introduce a proficiency test until 2012 (Kang, 2009). Even then, KICE may be unable to make the transition from a normative multiple choice exam to a proficiency test because of inadequate assessment training for teachers in such a short time frame. Negative backwash from high stakes tests can be expected to continue. Previous studies of high stakes tests have shown a high drop-out rates for students and teachers. Schools may record inaccurate grades while some students may learn how to fail due to the nature of normative testing (Amrein, 2002). There may also be no agreement on how to implement and assess a proficiency test as a quantitative assessment is viewed as more important and reliable than a qualitative one. To change from a teacher centered classroom to a student centered one may also prove problematic as their status would be threatened in a Confucian society. For the students, they learn at an early age what their ranking is in the classroom as a seemingly endless series of redundant tests are administered. Once pigeon holed, their situation may not change unless the teacher has aspirational expectations for them. Their friends at school soon become competitors as normative testing ensues. This may lead to fear and anxiety and be one of the contributors to the high suicide rate that exists in Korea today. Both the teachers and the students are subject to the heavy demands of middle school and high school curriculums. These curriculums are characterized by rote learning, a passive approach and multiple choice practice tests which make learning English or any other subject a burden (Ahn, 2003). For the family, they are limited in the amount of money that they can spend to privately educate their children for the highly competitive job market that awaits them. To solve this problem they may even separate with the mother and child sent abroad to ensure a proper education at a more affordable price. In a cultural sense, the issue of hierarchy and ranking people in society to promote social harmony may not be producing the desired results. These three groups may feel they are in a helpless situation with a cultural system that imposes educational and societal expectations that are beyond their grasp and leave them with a feeling of helplessness and inadequacy.

It should also be noted that the Ministry of Education has a poor track record at implementing and managing standardized tests. Ninety percent of the regional education offices misreported the results of new primary and secondary school standardized tests supposedly to avoid discrepancies in regional assessment (Kang, 2009). While this project has introduced surveys to compare the English language testing beliefs of students in Korea to better determine how they view language testing, it has not considered the inconsistencies or biases that exist in the education department. Whether this research project can address these different groups and create positive backwash is highly problematic as this research project only considers Korean university students and their perception of English language testing.

The cram schools that help perpetuate the endless series of testing are not likely to welcome a proficiency test because they lack skilled teachers and proper curriculums to get involved themselves. They may lobby the government to maintain the status quo or to prioritize testing of English such that more, not fewer, tests are created. Private education is so ingrained in Korea that parents are likely to continue to spend money so that their offspring obtain higher scores (Kang, 2009). The language test preparation industry is not likely to walk quietly into the night for fear of a proficiency exam.

In this research project, a paired criterion referenced test has been used as an alternative method of assessment. It was chosen because of the practicality of administering the test when large classes are involved. This testing method was also selected because of the positive effects that may be accrued from backwash, something the research believes is not occurring under the present testing method. However, this is not to say that other testing methods such as classroom based assessment, are not possible. The main issue is that evaluation in whatever form becomes more meaningful and relevant to the students so that they are able to benefit from positive backwash.

3. Methodology

3.1 Approach

Action research is appropriate in situations where changes in teaching methods and curriculum development are necessary. It directly addresses the problem of the division between theory and practice. It is research carried out by the participants and the researcher form the inside (Noffke, & Somekh, 2005).In this particular project it is hoped that a switch in language assessment will occur that better reflects the teaching methodology being employed.

The data will be collected formally through the use of surveys. The sampling will includes first year university students at a Korean university at the beginning of their first semester. They will be surveyed before and after a criterion referenced paired test. This test will assess the students on real life tasks similar to those they studied in class using a rating system based on four categories, namely fluency and coherency, lexical resource, grammatical range and accuracy and pronunciation. These students have not experienced a paired criterion referenced exam before. The students sampled are representative in terms of age, gender, and educational background of first year Korean university students. The sample size should be large enough to ensure the validity of the conclusions. The students sampled are from the researcher’s own classroom. The surveys will use a normative approach that allows the students an opportunity to state their level of agreement by indicating a position on a Likert scale (Horwitz, 1985). This scale will be modified to only four points (strongly disagree, disagree, agree and strongly agree). This will avoid ‘fence sitting’ on the issues contained in the sample instruments. This may be a flaw in the survey design as a greater number of options may result in more agreement among the people being surveyed (Brown, 2001). The responses will be quantified in the following manner: strongly disagree=1, disagree=2, agree =3, and strongly agree=4. The survey before the criterion referenced test will have 19 items (survey 1) and the survey after the test will have 24 items (survey 2).The survey items will include the subjective data of the students based on their beliefs of SL assessment. The survey items will include six categories of statements: the practicality of testing in Korea, reliability, validity, the impact that testing has (backwash), the authenticity of the test, and the interactiveness involved in the test. The survey items will vary on each survey. This has been done for the following reasons. First, asking the students about tasks they had to perform before a CR test would not be relevant as they had limited to no test experience with this test method. Second, the post CR test survey focused on asking the students about their perceptions of a CR test, again, something that they couldn’t answer in the initial survey. Third, after using the same language test method for 10 years, the researcher found it incumbent to use different survey items to better understand their perceptions and attitudes to both NR tests and CR tests and to tease out perceptions that were unique to either test method. However, it should be noted that 7 items are same on both surveys.

Variability will be measured using standard deviation and p-values. There will be three open-ended questions at the end of the survey to ask the respondents what they thought of the surveys. The students will be provided with a rubric so they are able to see how their mark was determined. The validity of the research will be enhanced and enriched by allowing them to compare the two testing methods after both surveys are completed.

3.2 Timeline

There will be 5 phases to the survey study:

Phase 1: The translated student survey is to be piloted on another teacher's students to make sure there is no misunderstanding of the survey items.

Phase 2: The student survey, written in Korean, will be administered by the researcher and possibly other teachers the first week of March, 2009.

Phase 3: The researcher will survey the students after a criterion referenced test so that they can compare the two distinct testing methods (June, 2009).

Phase 4: The data will be collected and coded using SPSS (Statistical Package for the Social Sciences). Statistical mean, standard deviation, and p-values will be tabulated.

Phase 5: The data will be interpreted and compared to SLA theories of learning and assessment methods.

3.3 Sampling

The sampling is critical to external validity or the extent to which finding can be generalized to people or situations other than those observed in the study (Pajares, 2004). In this particular study, the project is surveying first year university students. The students to be sampled are from the researcher's classes. They have just completed ten years of English language education in the public school system. Some of them have chosen a major field of study while others have yet to declare a major. The first sampling will occur at the beginning of the semester in class. The second sampling will occur after they have taken a CR test.

The items used in the survey are based on Kohonen's *Authentic assessment in affective foreign language education* (Kohonen, 1999). As far as the researcher knows, the instruments have not been used in a Korean context. The survey items have been grouped into language test qualities (Bachman & Palmer, 1996). For the pre-criterion referenced survey, the sample items are the following:

Practicality

2. Time limits do not allow me time to finish my test

6. Group testing is possible**.**

Interactiveness

1. When I take a test in English, I focus on only one correct answer.

12. Language testing is a relative competition (You win, I lose).

13. Current English tests forbid students to interact.

17. Students are compared with each other.

Reliability

3. Multiple choice exams have similar items to those I studied in class.

7. Multiple choice exams test only lower order knowledge.

10.The test offers the student a variety of different items.

Validity

4. I do not have to speak or write in an English test.

5. Tests emphasize what students cannot do.

9.The students should do a practice test before the actual test.

Impact

8. Test results reflect socio-economic status.

11. Current English tests are stressful.

15. Tests teach students why they fail.

16. It is embarrassing to speak in English.

18. I worry about making mistakes**.**

Authenticity

14. Test items simulate real-world tasks.

A post-criterion referenced survey will also be administered. This survey will compare the qualities of a criterion referenced test and with a norm referenced test. These items include the following:

Practicality

21. Group testing was possible.

Interactiveness

5.There was more than one correct answer

6. One word answers were not appropriate.

23. Current English tests forbid students to interact.

Reliability

1. The language test allowed me to complete a task

3. My test score reflected my abilities in English

Validity

12.The in-class tasks and the test tasks were similar.

13. The test tasks set standards of learning that were achievable.

14. The test tasks were appropriate given my ability in English.

Authenticity

2. Test items reflected real world situations.

6. One word answers were not appropriate.

12. The in-class tasks and the test tasks were similar.

Impact

4. The test motivated me to perform the task required.

7. The English language test was a relative competition (You win, I lose).

8. Students are evaluated on their ability to perform a task.

9. Students are compared with each other to determine a grade.

10. The test measured my progress in English.

11.By taking the test, I know what I can and cannot do in English

15. The test emphasized my strengths and progress in English

16. My confidence was enhance by taking this test

17. I was allowed to be successful.

18. The test was socio-economically fair.

19. The test allowed me to think and respond to a question.

20. Active awareness of learning was promoted

22. Fluency was more important than accuracy.

24. Tests teach students why they fail.

3.4 Data Collection and Analysis.

The surveys are to be collected from the students before and after a criterion referenced test. The first survey is to be conducted right after the first class at the beginning of the semester. The second survey is to be conducted right after a criterion referenced test. Quantitative values are to be assigned to the Likert scale to analyze the data. The values that will be calculated include mean scores, standard deviation and percentage agreement for all the items. Comparisons of mean scores and standard deviation will be analyzed when the survey items are the same. A one-tailed t-test will also be conducted on same survey items that have been included in both surveys.

4. Project Results and Discussion

4.1 Impact

Statistical Table of Results

Table 1. Pre Criterion Referenced TEST Student Survey 1.

Survey Item	Mean	% agree	St. Dev.
Item 11. Current English tests are stressful	3.03	76	.486
Item 12. The test was a relative competition (you win, I lose).*	3.03	76	.518
Item 18. I worry about making mistakes.	2.95	74	.688
Item 17. Students were compared with each other.*	2.92	73	.531
Item 15. Tests teach students why they fail.*	2.74	69	.549
Item 16. It is embarrassing to speak in English.	2.74	69	.677
Item 8. English test results reflect socio-economic status.*	2.31	58	.731

Statistical Table of Results

Table 2. Post Criterion Referenced Test Student Survey 2.

Survey Item	Mean	% agree	St. Dev
Item 20. Active awareness of learning was promoted.	3.22	81	.591
Item 15.The test emphasized my strengths and progress in English.	3.14	79	.593
Item 11. I was able to determine what I can and cannot do in English.	3.11	78	.575
Item 19. The test allowed me to think and respond to a question.	3.08	77	.500
Item 4. The test motivated me to perform the task required.	3.03	76	.506
Item 7. The test was a relative competition (you win, I lose) .*	2.72	68	.508
Item 9. Students were compared with each other.*	2.67	66	.717
Item 8. I was evaluated on my ability to perform a task	2.61	65	.506
Item 24. Tests teach students why they fail.*	2.58	65	.768
Item 18. English test results reflect socio-economic status.*	2.53	63	.810

*same item

Impact is a test quality that affects society, the educational system, and the individuals directly affected by the test. Thus impact operates on a macro or societal level and a micro or individual level. Both the act of taking a test and the use of test results imply values and goals that have consequences for the student, the teacher and society (Bachmann and Palmer, 1996). This is because of an impact trait called backwash, or the beneficial or harmful effect of testing on teaching and learning (Hughes, 1989). In this section, the survey results will be analyzed and discussed in terms of both positive and negative backwash.

The normative assessment employed in Korea reflects the use the education authorities have for the test. This use is characterized by an examination approach which tests language skills in isolation using accuracy in language use to ensure both objectivity and reliability in scoring. While this may be a valid method in certain circumstances, particularly for selection purposes, in Korea it excludes the macro kills of writing and speaking. The extensive use of multiple choice exams complements this structural view as they severely restrict what can be tested. This method of testing, which is based on language form recognition, may also trap students into making incorrect responses through the use of distracters. It has a profound impact on the teachers as they must teach to the test and the students who must absorb vocabulary and grammatical points. This kind of testing reflects the importance of the high-stakes test (CSAT) that every high school student must pass for university entrance.

Passive learning complements this approach as there is little opportunity for students to absorb new information or employ their cognitive abilities to practice using this information in a communicative way. This learning characteristic seems to be reflected in survey 1, items 11, 16, and 18. Before the CR test, the students were embarrassed to speak in English (item 16, 69% agreed). Their mild agreement with this item may be interpreted literally but also pragmatically as this is probably the first time they have ever been surveyed about their English education and their own subjective views. In item 18 (74% agreed), they worry about making mistakes, which is characteristic of passive learning in the Korean classroom as students are generally seeking an answer and not a conversation. In item 11, (76% agreed), they view current English tests as stressful. This stress may be a result of the competitive environment that the students have to deal with on an ongoing basis if they are to be successful (selected) for university admission. While there is no relationship between passive learning and stress, their positive response on this item would seem to indicate that competition and the inherent stress it produces dominate their approach to learning to the detriment of an interactive approach to learning a language with their peers.

When surveyed after a CR test at university, the students suggest that they are more active learners. Their 81% agreement on survey item 20 of table 2 that “active awareness of learning was promoted,” seems to indicate that a change in their learning styles is occurring. This response is complemented by survey item 19 of table 2. “The test allowed me to think and respond to a question,” had 77% agreement. This seems like a relatively short period of time to make such a profound change in learning styles as only a year has passed since the majority of students left high school. It would also seem that the impact that one shot tests is quickly forgotten by students. Perhaps the students didn’t like the aim of one-shot high stakes tests, which compared them with the performance of other students for selection purposes. Instead, they seem to prefer a CR test that doesn’t emphasize a correct answer but whether they were able to accomplish certain language tasks. This apparent change from passive to active learning may also reflect the fact that the goals of the students are changing. Passing a one-shot exam is no longer the issue. Instead, improving English communication skills to enter the job market may be becoming their priority.

Because high school students must prepare to pass a one-shot university entrance exam, they are motivated extrinsically to pass a test as opposed to intrinsically learning a language. Thus their motivation is utilitarian in nature. If the testing method was changed, perhaps a motivational shift would occur. By using other testing methods such as a CR test or direct testing, the students would have a clearer picture of what they have to achieve. They realize that by performing the test at a criterion level, they will be successful regardless of the outcome of other students. This will have a pronounced effect on motivation and appears to be reflected in the survey items. In survey 2, item 4, their 76% agreement indicated that “the students were motivated to perform a task.” That a CR test “emphasized their strengths and progress in English” (survey 2, item 15) had 79% agreement. This seems to indicate that the students have a positive attitude towards language learning**.** The use of meaningful standards helped students “determine what he can and cannot do in English” (survey 2, item 11), and received 78% agreement. This suggests that students can measure their abilities in relation to criteria they are familiar with (Hughes, 1989). That the test task “allowed me to think and respond to a question” (survey 2, item 19) received 77% approval, again suggesting positive backwash. That you can extrapolate from this that a motivational change is underway from passive to active learner, from extrinsic to intrinsic motivation, requires more research. These results do suggest, however, that motivation is related to cognitive constructs, self-efficacy in a language task, and language learners constructions of self confidence that positive backwash can nurture and promote (Bachmann,1990). It becomes incumbent on the teacher to test what the student can do in a language rather than what they cannot do and to make the students aware that progress is being made. Both direct tests and CR tests can serve as devices through which positive backwash is obtained. By evaluating a student in a qualitative manner, a teacher is able to praise the strengths as well as constructively criticize weaknesses. This can be accomplished by having the test at the penultimate session and the evaluation at the last session. To enhance positive backwash, the students may even discuss the evaluation judgments. Therefore, it is possible that motivation is promoted with this approach, whether it be intrinsic or extrinsic in nature.

Norm referenced testing, which is common in Korea, can be characterized as being developed independently of any particular course of instruction. This type of testing would have previously been administered to a large sample of students that form the target population to enable the test user to make ‘normative’ interpretations of the test results. The test results of the sampled group serve as a reference point for interpreting the performance of other students who take the test. If this type of test is properly designed, a score will be typically distributed on a bell-shaped curve. This implies that these tests are only valid with the population on which they have been normed (Bachmann, 1990).

4.2

Survey Item: The test was a relative competition. (You win, I lose)

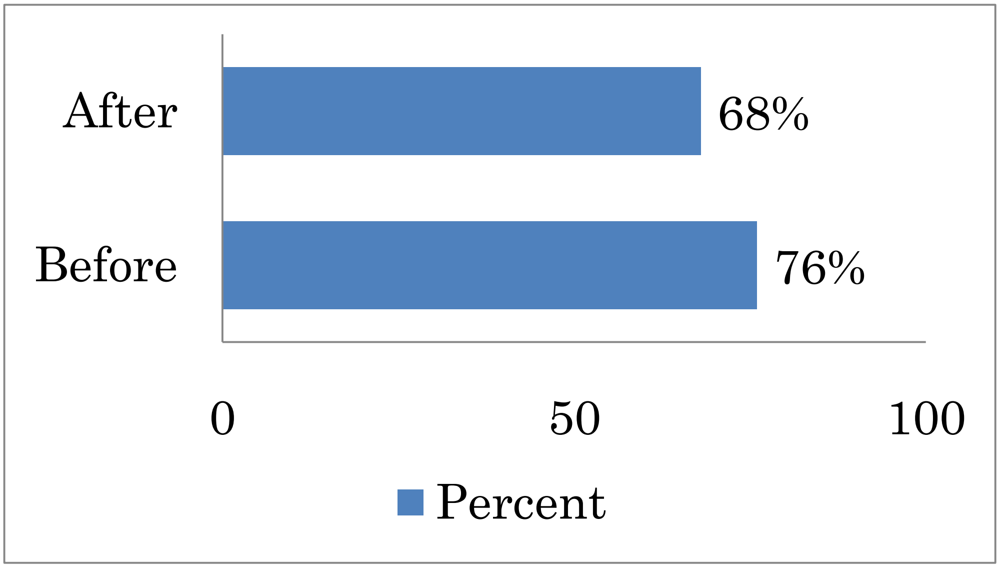

Figure 1. The students perception of test impact before and after a CR test.

Norm referenced testing might account for the survey results in figure 1. That “the test was a relative competition (You win, I lose)” had 76% agreement for the students before a CR test and 68% afterwards. The higher agreement about the competitive nature of a standardized test may reflect the fact that the students scores are being compared one against the other. They disagreed 8% more on the survey item after a CR test. They may realize that they were scored on their performance to do a task and not against each other. A CR test has tasks that students must respond to as they are being evaluated on their ability to perform them. It may imply that the students became motivated to successfully respond to these tasks and in the process achieve the standards set out in the course. This is in direct opposition to a NR test that tests students against each other and may not use a meaningful standard (Hughes, 1989).

4.3

Survey Item: Students were compared with each other

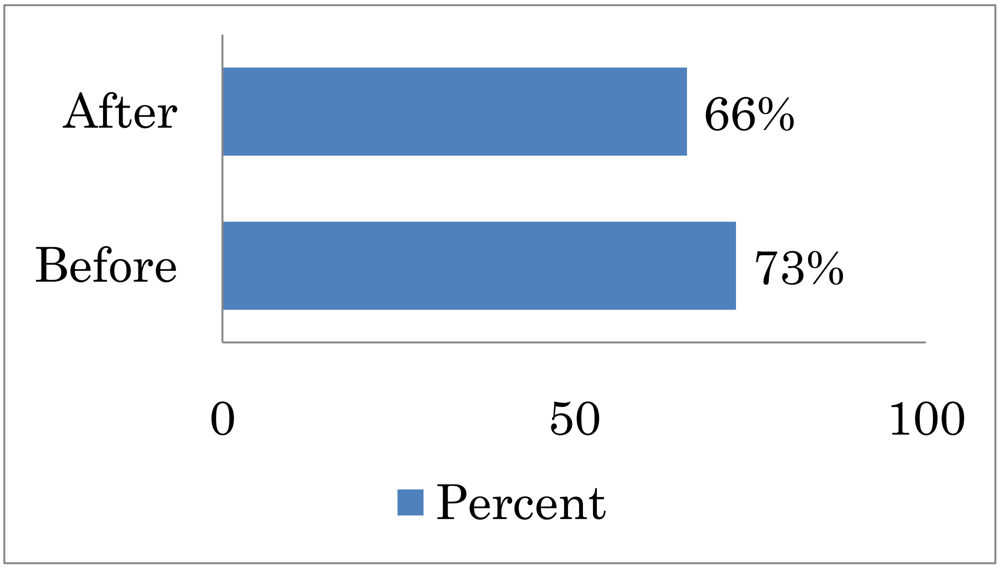

Figure 2. The students perception of test impact before and after a CR test.

This theme appears to be corroborated in figure 2. The survey item states that “students were compared with each other.” Before the CR test there was 73% agreement while agreement dropped to 66% after the CR test. The 7% decrease would seem to indicate that the students realize that they are being tested on their performance and not against one another. The reason the difference is only 7% may have to do with a grading curve. The students realize that they have to be ranked against their peers to obtain a final grade. Thus the test is essentially normative, even though the testing method was criterion referenced. This becomes a serious problem in classes where students are simply lumped together without a placement test as the results may not be valid. It could mean that if most of the students in one class at a university did poorly on the test relative to a large university population and were then curved up, they could receive final scores that were higher than students with higher earned scores that were curved down.

The method of testing and its measurement using a grading curve may confuse the students as survey 2, items 4 and 8 would seem to indicate. Item 4 states “the test motivated me to perform a task” and had 76% agreement. Item 8 states that “I was evaluated on my ability to perform a task” and had a 65% agreement. The greater disagreement (11%) with the later survey item may be because they realize that they are to be ultimately ranked and compared to their peers and not against their ability to perform a task. This creates negative backwash as the effect of the test is mitigated by the final test score even though the students fully knew and understood what was required of them.

4.4

Survey Item: Tests teach students why they fail.

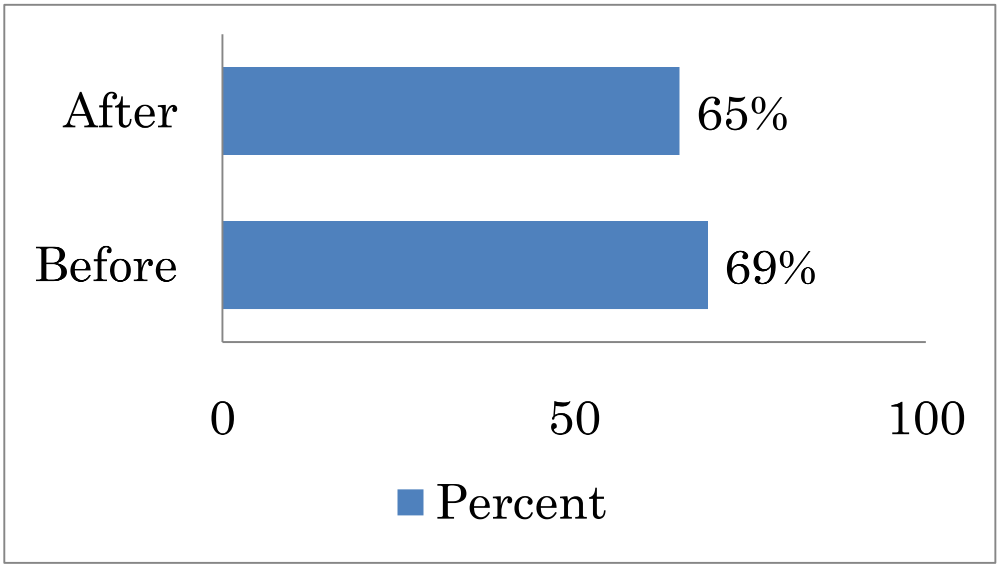

Figure 3. The students perception of test impact before and after a CR test.

The students are essentially neutral on survey item 24 which states that tests teach students why they fail (figure 3). Their mildly positive response may have to due to the CSAT, which most of them have taken in the last 6 months. The content of this final achievement test must be related to the syllabus but the percentage of the material tested may vary widely. In the case of Korea, there is currently no written or oral examination. This is probably the most important reason why the students offered the same opinion before and after the CR test. Their 69% and 65% agreement suggests that neither the students nor teachers or both of them don't really know what the purpose of a test is outside of a tool for ranking and selecting students. If only certain parts of a syllabus (reading and listening) are used, it is also logical to conclude that in a CLT curriculum, the objectives of the course are not being met. When this occurs the test results are of limited validity and application to other language testing models.

This survey has also not covered the important role that the education authorities have on test impact. The use of test scores implies that there are societal values and goals that have consequences at both the micro and macro level and that KICE is empowered to make decisions in this regard. In April, 2009, KICE, for the first time, released the results of the CSAT on a school by school basis. By doing so, they were effectively ranking schools and putting them into competition to raise scores. This action may encourage students to attend only select schools while possibly turning them into test preparation facilities at the expense of education. This decision would seem to be at odds with the government's policy of choosing university students based on talent and potential (Kang, 2009).

4.5 Reliability

Statistical Table of Results

Table 3. Pre Criterion Referenced Test Student Survey 1

Survey Item	Mean	% agree	St. Dev.
Item 12. The test was a relative competition (you win, I lose).	3.03	76	.518
Item 18. I worry about making mistakes.	2.95	74	.686
Item 17. Students were compared with each other.	2.92	73	.531
Item 10. The test offers a variety of different items.	2.72	68	.510
Item 14. Test items reflected real-world situations.*	2.54	64	.643
Item 8. English test results reflect socio-economic status.*	2.31	58	.731
Item 3. Multiple choice exams have similar items to those I studied in class.	2.13	53	.362

Statistical Table of Results

Table 4. Post Criterion Referenced Test Student Survey 2

Survey Item	Mean	% agree	St. Dev
Item 11. I was able to determine what I can and cannot do in English	3.11	78	.575
*Item 1***.** The test allowed me to be involved in a variety of tasks.	3.08	77	.439
Item 2. The test items reflected real world situations.*	3.08	77	.500
Item 14. The test tasks were appropriate given my ability in English.	2.99	75	.465
Item 13. The test tasks set standards of learning that were achievable	2.94	74	.630
Item 18. English test results reflect socio-economic status.*	2.53	64	.810

*same items

Reliability is a language test quality that is concerned with consistency in scoring. While you can never have complete consistency in scoring, measurement errors in assessment are minimized so that the results are dependable. Sources of errors in language testing can vary from the test method, to personal attributes, to communicative language ability, to random factors. Thus, any investigation of reliability has to consider how much of an individual's test performance is due to measurement error or how much error is due to other factors. (Bachmann, 1990).

To guarantee reliability on a language test, a large number of items usually ensures a greater range of scores that effectively separates candidates. If this same test were administered on a number of occasions and the same result obtained, reliability would be validated (Henning, 1987). In survey 1, item 10, "the test offers a variety of different items," the students before a criterion reference test mildly agreed (68%). The reason they mildly agreed may reflect negative backwash they have experienced as a result of norm referenced tests. The students realize that their performance has been compared to other students and not against what they are capable of doing in a second language. They may also understand the criteria for correctness is limited in a multiple choice exam as usually only one correct answer is possible. If there is no penalty for guessing, then they can simply engage in an elimination process among a few items to arrive at the correct answer. Because this method of testing measures form recognition and has the same format, the students find this method of testing limited in nature. When surveyed after a criterion referenced test on a similar survey item (survey 2, item 1), the students agreed (77%) that "the test allowed me to be involved in a variety of tasks." This change in test format allows a student to construct an answer or complete a task as opposed to select an answer. These tasks are usually context-embedded so the student is familiar with the task, the participants involved, the situation and the probable outcome. Thus, students can recognize what the task is testing as long as it is presented to them a meaningful way.

Korea is a very test intensive country and due to both practicality and objectivity relies heavily on NR tests for selection purposes. But because of their limitations, they are not a comprehensive measurement of language mastery. In many instances these tests are more sensitive to inter individual differences so their estimates of reliability with CR tests are limited (Heaton, 1987). After their CR test, the students seem to suggest (74%) that “the test task set standards that were achievable” (survey 2, item 13). In other words, the test was reliable as the rationale for the test, its specifications and the sample items made the student aware of what was required to be successful. Furthermore, a CR test can provide positive backwash as it allows students to interpret a test score with reference to a criterion level of ability and their mastery of it. This is not the case in NR tests as they only compare students against each other. The students concur that “they were compared with each other” (survey 1, item 17, 73%) and that “the test was a relative competition, you win, I lose” (survey 1, item 12, 76%).

The large number of test items in a typical NR test and objective marking usually make the test reliable. But this type of objective test does not allow the students to respond more personally to a task as one clear unambiguous answer is presented to them (Mangubhai, 2006). In a CR test, the sample items tested must be representative of the domain you want to make inferences about. If the items are consistent, then the observed score should be a reliable indicator of the domain you want to make inferences about. The students agree (78%) with this principle of subjective scoring. That “I was able to determine what I can and cannot do in English” (survey 2, item 11) implies that not only the examiner but also the student can determine ability, something that does not occur in NR testing. Feedback may become reciprocal as the examiner provides guidance regarding their performance and possible problem areas while the students provide feedback to the examiner on the limitations of the test. Indeed, after their CR test, they agree (78%) that “I was able to determine what I can and cannot do in English”(survey 2, item 11).

It becomes easier to revise a CR test because of the feedback it generates. Teachers are more informed about the entry and exit levels of their students and whether the objectives of the course were realized. Students relate their needs to the teacher so that the objectives of the course can be adjusted to tease out goals and objectives that may have been redundant while introducing new ones that are more appropriate to their needs. The teacher can adjust the course content and revise the CR test accordingly. Finally, a CR test would involve evaluation, whether it be formative or summative. This process is meant to serve the students so that that their language education becomes a rich and meaningful experience. In stark contrast, a NR language test in Korea is only providing the students with a numerical evaluation and not providing them with the same kind of learning experience. In addition, the test writers may not realize that the information gathered from the test is essential not only to measure performance but to analyze and revise so that future tests reflect the needs of the students.

4. 6

Survey Item: English test results reflect socio-economic status

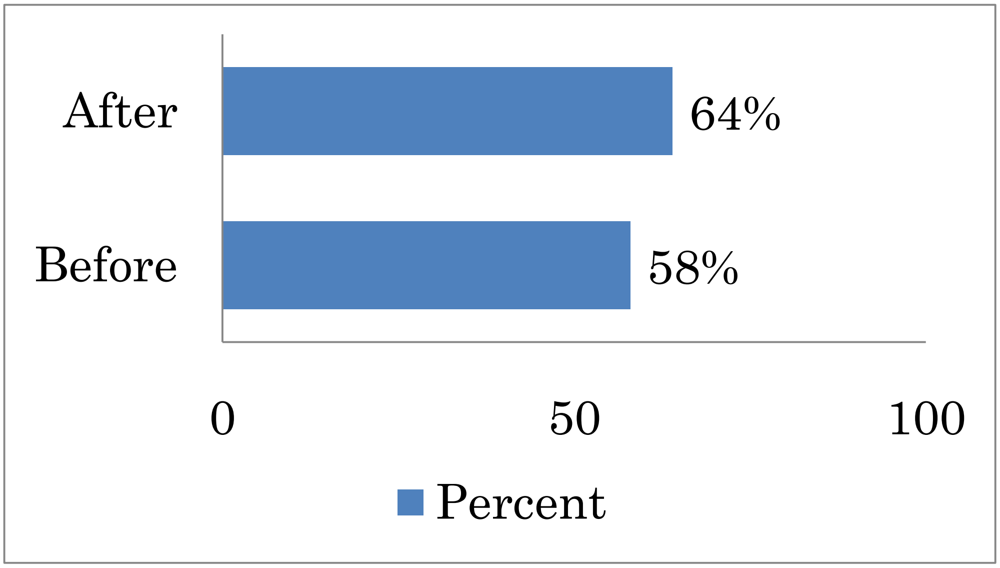

Figure 4. The students perception of private education before and after a CR test.

.

Test performance may be influenced by other attributes such as real world knowledge and socio-economic background. In figure 4 the students agreed more that “English test results reflect socio-economic status” after taking a CR test. This may be interpreted to mean that they all don’t come from similar family backgrounds and have had the same exposure to the thriving private education market in Korea. Thus, when they take a proficiency test they are not all equal in their exposure to English language education. The small 6% difference could also mean that their private education has had limited influence on their ability to perform well on a proficiency test. This is more likely as the majority of private institutes are not skilled in speaking and writing training to prepare students for standardized proficiency exams such as IELTS.

While a decrease in the number of students enrolled in private education is occurring, this may be because of a slow down in the Korean economy and a decreasing birth rate and not socio-economic status. Indeed, spending on private education is actually increasing in households whose earnings are between 5 and 7 million won (Hartman, 2009). That, coupled with the recent release of CSAT scores by region and school, may fan the flames of increased private spending and transfer of students to high schools and areas that scored significantly higher on the CSAT. It may act as a catalyst to increase private education spending by parents (Kang, 2009). In terms of perception, this 6% difference could decrease once CR tests are introduced and adopted as students will realize that it is a criterion they are trying to achieve and that their score is a measurement of mastery of that criteria. They may also realize that the decision rendered over mastery is not influenced by the scores of other candidates. In fairness to the students, this survey item should be answered by the parents to determine their views on private education.

CR tests are designed so that test score interpretation can be representative of criterion ability. In many instances the test scores are similar as the students develop uniform ability as the course and its objectives are met. This may question the reliability of CR test scores in terms of consistency, dependability and stability. To overcome this potential problem, the test maker must insure that the test items are representative of the domain to which inferences are to be made. If the test items are highly consistent and if the test items are equivalent, then the resulting observed scores should be reliable indicators of domain scores (Bachmann,1999). The students concur (75%) with this approach to score interpretation. “The test tasks were appropriate given my ability in English” (survey 2, item 14) implies that the items used on the test were reliable indicators of their ability in the domain being tested. Their ability to make inferences about a CR test contrasts with their interpretation of a NR test. They disagree (53%) that “multiple choice exams have similar items to those I studied in class” (survey 1, item 3).

While it is generally agreed that this type of test is subject to guessing and elimination of distracters, their response seems to indicate that they were not being tested on the skills they had acquired. This could mean that Korean NR tests have an overabundance of test items that try to segregate students, but at the same time, infringe on content validity.

4.7 Practicality

Statistical Table of Results

Table 5. Group testing comparison before and after a criterion referenced test

	Before (Survey 1)			After (Survey 2)		
Survey Item	Mean	% Agree	St. Dev.	Mean	%Agree	St. Dev.
Group testing was possible.	2.67	67	.577	3.00	75	.586

Practicality refers to the ways the test will be implemented, what resources are available and whether they are available to administer the test. These resources may be human, material or time related. The demands of the test specifications can only be realized if these resources available are adequate. If this is not possible, the test will not be practical (Bachmann and Palmer, 1996).

Any shift from a multiple choice format to a proficiency model in language testing would require development time as the current testing procedure for high school students does not have a speaking or writing component. The development time required would be substantial as design, writing, administration, scoring and analysis would all have to be carried out prior to the actual implementation of a CR test. Specifications for the test would have to be written at the outset and include the tasks the students would have to carry out. The tasks would have to satisfy the objectives of the curriculum. A critical level of performance would have to be established and determined using descriptive bands so that speaking and writing objectives are assessed correctly. These bands must then be calibrated by using sample performances and then rated by trained examiners in the field to ensure reliability. All high stakes tests would have to go through this process to ensure the validity of the results (Hughes, 1989). As a consequence, the Eiken Test in Practical English Proficiency is being considered for implementation in 2013 by the Korean Ministry of Education, Science and Technology. While it may be deemed a practical move to save time, it may not create positive backwash as it may set off a new series of private university admission exams that question the credibility of a test that evaluates students by level and not by scores (Kang, 2008).

4.8

Survey Item: Group testing was possible

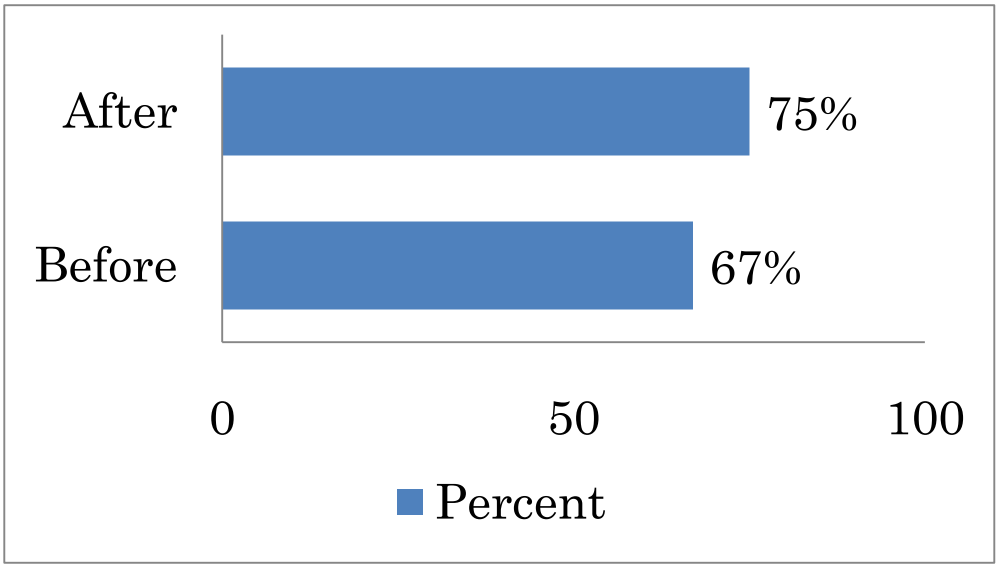

Figure 5. The students perception of the practicality of group testing.

While a test should be easy and inexpensive to construct, administer, score and interrupt, this is not the case with proficiency tests. The time required to administer a new proficiency test would be substantial. A possible solution to the time issue is the use of group testing. On the survey question "group testing is possible" (figure 5), the students agreed 8% more with the statement after taking a CR test. This change in response may parallel their traditional beliefs of language testing with its emphasis on accuracy and one word answers with a test task that invites them to interact with their partner. Also, because they were able to cooperate with their partner during the test and not have to deal with the intimidation of speaking directly to the examiner, they were better able to engage in the task. But, their more positive response probably had to do with the relative ease of performing a simple real world task they were familiar with. This feature of testing does not always occur in a NR test, which is usually testing for accuracy in a more competitive setting. While there is no direct survey item on promoting a more cooperative approach to learning and testing, by giving the students a task in which they have to cooperate with one another, interaction is promoted. This interaction seems to promote a richer and more varied use of language than if the candidates had been tested alone (Taylor, 1999). However, it should be noted that group or pair testing is not without its limitations. If one candidate dominates the test task, a true interpretation of another student's ability may be compromised and threaten the validity of the test result (Hughes, 2003). In the latest suite of Cambridge ESOL proficiency tests, this is mitigated to some degree by having the candidates exchange roles (Cambridge ESOL, 2009).

While contemporary methodology such as authentic pedagogy would include student empowerment to design language tests, a top down hierarchical approach to education and testing dominates the Korean cultural landscape (Kim 2004). Ironically, the students were not surveyed on this issue of empowerment as they are the very people to benefit from teacher-student interactive approach to language and testing methodology. In a practical sense, it almost certainly would benefit both students and teachers once the students determine the objectives of the course and the skills that are to be tested. Thus, a natural progression from test taking, to self assessment, to designing tests can occur while taking into consideration the limitations of the teacher's resources and time. Whether this method of testing becomes popular in a traditional society remains problematic.

4.9

Interactiveness

Statistical Table of Results

Table 6. Pre Criterion Referenced Test Student Survey 1

Survey Item	Mean	% agree	St. Dev.
Item 13. Current English tests forbid students to interact. *	2.51	63	.556

Statistical Table of Results

Table 7. Post Criterion Referenced Test Student Survey 2

Survey Item	Mean	% agree	St. Dev
*Item 1***.** The test allowed me to be involved in a variety of tasks.	3.08	77	.439
Item 5. There was more than one way to answer a question.	3.06	77	.583
Item 3. My test score reflected my abilities in English.	2.94	74	.538
Item 13. The test tasks set standards of learning that were achievable	2.94	74	.630
Item 23. Current English tests forbid students to interact. *	2.25	56	.792

* same item .

Interactiveness is a relative language test quality that involves the extent and type of an individual's test taker abilities that are used in accomplishing a test task. These abilities include language knowledge, strategic competence and metacognitive strategies, topical knowledge and affective schemeta. Interactiveness forms a link with language constructs such as competence in the validation of language tests (Brown, 2004).

4.10

Survey Item: Current English language tests forbid students to interact.

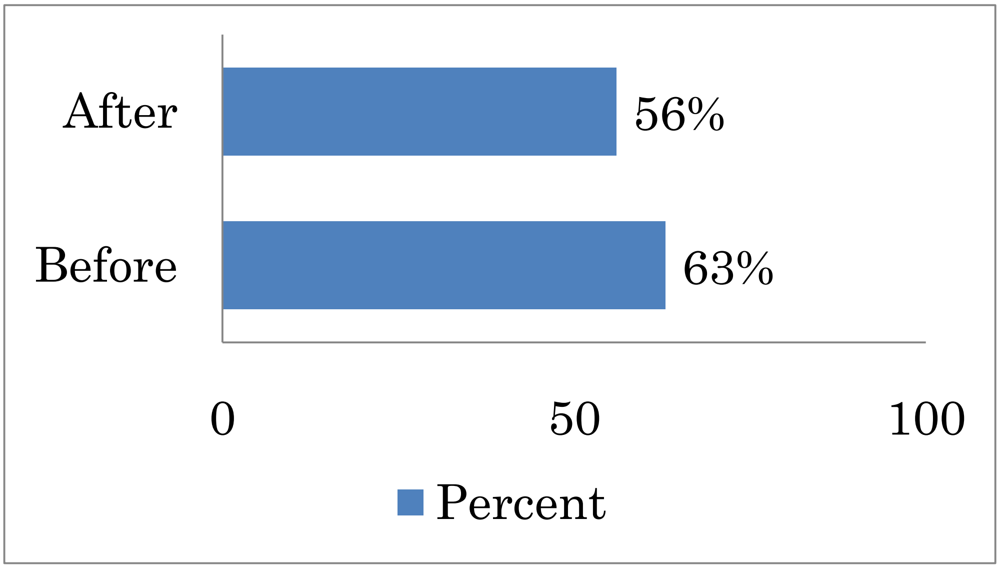

Figure 6. The students perception of interactivity before and after a CR test.

The students' interpretation of the word interact is probably more literal and less academic in figure 6, as the students barely agree that "current English language tests forbid students to interact." This may reflect the influence of multiple-choice exams that exclude interaction and focus on testing lower order knowledge. NR tests may also circumscribe the area of language knowledge being tested. As a result, the extent and involvement of a test taker's abilities are limited. This would marginalize construct validity as any inferences made from the test would be a function of the abilities tested (Bachmann and Palmer, 1996). It is interesting to note that 7 % of the students no longer thought that tests forbid students to interact after the CR test while half of them appear not to have changed their mind.

The students had a different view of interactiveness when they were interviewed after a criterion referenced test. In survey 2, item 1, the students agree (77%) that "the test allowed me to be involved in a variety of tasks." This suggests that "there was more than one way to answer a question" (survey 2, item 5, 77%). The relative degree of interactiveness would seem to be substantially higher after this method of testing. This, in turn, would have a positive effect on construct validity as more language abilities are being employed by the student. Thus, inferences made from assessment become more meaningful than a multiple-choice exam that does not tell you whether a student can function in a foreign language (Hughes, 2003).

The students seem to better interact with the test task as "there was more than one way to answer a question" (survey 2, item 5, 77%). This may imply that they were using different abilities such as the context of the situation, phonology and/or grammatical competence to formulate a response. Their interactiveness with the test task was noted in their self-assessment. They agreed (74%) that "my test scores reflected my abilities in English" (survey, 2, item 3). This may reflect the fact that a real world test task influenced their ability to determine how well they answered the question. The use of real world or familiar topics seems to have an added advantage of acting as a catalyst for self assessment. This may minimize negative test bias and maximize the positive backwash of testing procedures (Bachmann, 1990).

Besides a degree of mastery of a content domain, the students also indicated that a level of ability may have been achieved. That "the test tasks set standards of learning that were achievable" received a positive response from the students (survey 2, item 13, 74%). Interactiveness will vary from student to student due to its relative nature. When students are tested, their processing strategies are determined by the test task and their innate language abilities. This can make it difficult to make inferences about what and how much of a language ability was used. Because interactiveness and construct validity are linked, this implies that the inferences made about language abilities may be suspect and the process of construct validation tenuous. However, the students seem to be validating the criterion referenced test they took by responding positively.

4.11 Validity

Statistical Table of Results

Table 8. Pre Criterion Referenced Test Student Survey

Survey Item	Mean	% agree	St. Dev.
Item 9. Students should do a practice test before the actual test	2.69	67	.614
Item 3. Multiple choice exams have similar items to those I studied in class.	2.13	53	.362

Statistical Table of Results

Table 9. Post Criterion Referenced Test Student Survey.

Survey Item	Mean	% agree	St. Dev
Item 15. The test emphasized my strengths and progress in English.	3.14	79	.593
Item 11. I was able to determine what I can and cannot do in English.	3.11	78	.575
Item 14. The test tasks were appropriate given my ability in English.	2.99	75	.465

While reliability is an agreement between efforts to measure a trait through two similar methods, validity is an agreement between attempts to measure the same trait through two different methods (Bachmann, 1990). With this test quality, the test scores must be interpreted correctly and demonstrate what they claim to measure so that decisions made are not arbitrary but based on a value system that justifies them (Manguhbai, 2006).

Sometimes in high stakes achievement tests, the test may wrongfully require grammar and vocabulary that the students were not aware of. This type of test lacks content validity (Henning, 1987). This seems to be the perception of the students to survey 1, item 3, which states “multiple-choice exams have similar items to those I studied in class.” The students 53% agreement seems to indicate a content related problem for a number of reasons. Perhaps they have always done poorly using this method of testing and are making a subjective judgment or perhaps it reflects the distinct advantage that criterion referenced tests have as content validity is usually assured. Indeed, on survey 2, item 14, the students agree (75%) that “the tasks were appropriate given my ability in English.” Their positive response seems to imply that both the content and the method of testing were valid considering the objectives of the course. By basing a test on objectives rather than detailed vocabulary and grammatical content, a better understanding of what has been achieved can be realized (Hughes, 2003).

To insure content validity, a practice test is warranted. The students mildly agree (67%) that "students should do a practice test before the actual test" (survey 1, item 9). They may be suggesting that some students can develop a better ability to answer unfamiliar questions that are not content embedded. This may be due to test wiseness or the ability to perform better or worse on a multiple-choice exam depending on the number of times the test is taken. The observed score may overestimate the true score in the case of someone who has taken the test repeatedly as repeat test takers tend to score higher on multiple-choice exams. This is not necessarily a threat to validity in a NR test as the purpose is to discriminate between individuals within a specific group. A selection test like the CSAT would develop a measurement scale that results in a score distribution that recognizes this purpose and therefore is reliable and valid. This may be why the student only agree (53%) with the statement that "multiple choice exams have similar item to those I studied in class".

A CR test is quite different as content or domain mastery determines the individual score of a candidate. To ensure validity, the specifications and objectives of the course and the test task (content) would all have to complement each other. The issue with CR test validity is the inference that can be made that the observed score is a reliable indicator of domain mastery. The students touch on the issue of content and domain mastery after a CR test. They agree (75%) that "the test tasks were appropriate given my abilities in English" (survey 2, item 14). "I was able to determine what I can and cannot do in English" (survey 2, item 11, 78%) and "the test emphasized my strengths and progress in English (survey 2, item 15, 79%). They may be implying a CR test allows them to gauge where they are in terms of both ability and domain mastery, something that wasn't possible after a NR test. This attribute of a CR test would be particularly beneficial in a country like Korea where over testing borders on the irrational and seems to be related to rank and hierarchy (Stevens, 2009). Between feedback from teachers and the self-assessment a CR test seems to provide, a Korean student could, at last, determine to some degree where they are in their language development.

4.12 Authenticity

Statistical Table of Results

Table 11. Pre Criterion Referenced Test Student Survey 1

Survey Item	Mean	% agree	St. Dev.
Item 14. Test items reflected real-world situations.*	2.54	64	.643
Item 3. Multiple choice exams have similar items to those I studied in class.	2.13	53	.362

Statistical Table of Results

Table 12. Post Criterion Referenced Test Student Survey 2

Survey Item	Mean	% agree	St. Dev
Item 2. The test items reflected real world situations.*	3.08	77	.500
Item 10. The test measured my progress in English.	3.03	76	.560
Item 14. The test tasks were appropriate given my ability in English.	2.99	75	.465
Item 3. My test score reflected my abilities in English.	2.94	74	.538

*same item

Authenticity is the degree of correspondence of a given language test task to the features of target language usage that are being assessed. It is linked to construct validity as it relates the test task to the domain of generalization from which you want score interpretations to make inferences about. Authenticity is also important because of its effect on the test taker's perception of the test and its relevance. The material used in the test task should be as natural as possible, contextualized, meaningful, and closely approximate real-world tasks. Using relevant material usually helps promote a positive response and perception from the test taker. (Bachmann and Palmer, 1996).

4.13

Survey Item: The test items reflected real world tasks

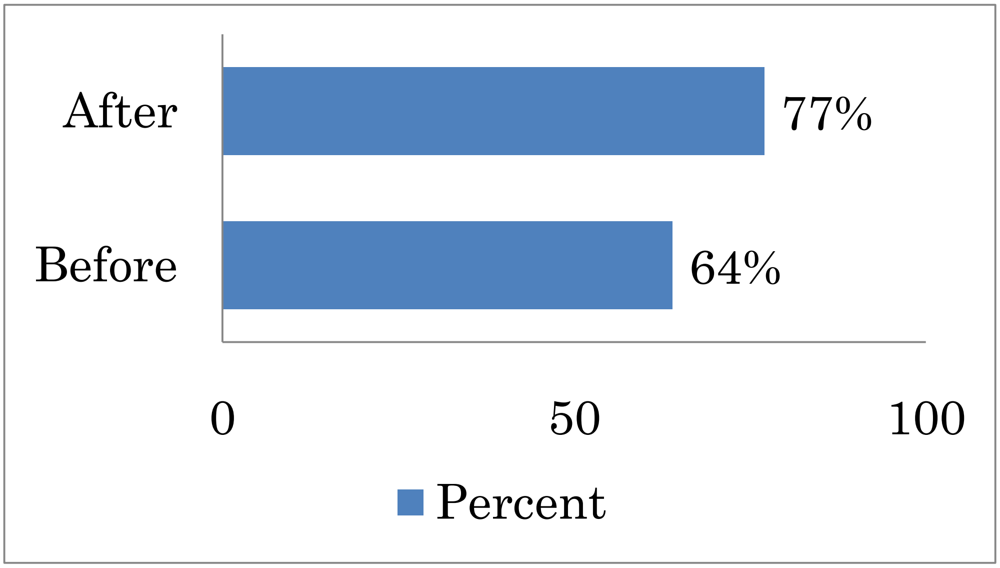

Figure 7. The students perception of authenticity before and after a CR test.

It is not surprising to find that before a CR test students slightly agree (64%) that "test items reflected real world tasks" (Figure 7). After a CR test, the percentage of students increased dramatically on this survey item as they performed test tasks that they were familiar with. These test tasks varied from asking for and giving directions to shopping for electronic devices. The 13% increase may have a positive effect on the interpretation of scores as a candidate's prior knowledge of the situation may act as a self assessment tool in determining language ability. On the other hand, in a multiple–choice exam, a student's capability to measure ability is compromised by factors such as guessing and adeptness at taking multiple choice exams.

The testing method that they are most familiar with is norm-referenced, which discriminates between candidates. As such, the test items may vary widely as this approach supports distinctions among students. The students disagree (53%) that "multiple choice exams have similar items to those I studied in class," (survey 1, item 3). Their disagreement supports the premise that the content and thus the authenticity of a

CLT curriculum have been violated when they were tested. On the other hand, a CR test has specific constructs that use real world language it wants to test and make inferences about.

The issues with authenticity are twofold: whether the test task requires the test taker to perform a behavior that simulates what would occur in real life and the relevance of target language usage to a broader domain of language use so that the score can be more widely interpreted and inferences made (Bachmann, 1999). Because of the complexities of real-life language use, the relationship between the test task and real-life use can become tenuous. This is particularly apparent when a test taker has had a sheltered, academic upbringing. The limitation of the test task response in such a situation may have a harmful effect on construct validity as the information gathered is insufficient to make inferences about (Bachmann, 1999). It would seem that a real-life approach to testing (authenticity) would have to consider a variety of different topics to obtain more information to improve diversity and thus construct validity. This may prove impractical when time constraints limit testing. There seems to be a need in this area to improve tests as current proficiency exams such as IELTS can take up to 14 minutes (IELTS, 2009).

The perception of the test by the test taker is an important feature of authenticity. If the test is perceived as being more relevant by the candidate, a more positive response and hence a better performance by the candidate can be expected (Bachmann and Palmer, 1996). A positive feeling about criterion referenced tests was indicated in several survey items. The students agreed 76% that “the test measured my progress in English,”(survey 2, item 10), agreed 74% that “the test score reflected my abilities in English” (survey 2, item 3) and agreed 75% that “ the test tasks were appropriate given my ability in English” (survey 2, item 14). This seems to indicate that the test was relevant to the students even though this was probably the first time they had taken a test in English that was interactive. This contrasts with a multiple-choice exam that is non-reciprocal as the candidate is totally unaware of the effect of their response. Given the fact that certain candidates respond differently, their responses may vary and thus change the test language. This feature may force the examiner to adapt to the candidate’s perception of the task and interact accordingly. This flexible feature of a CR test enhances both authenticity and positive backwash.

Authenticity is related to content validity because of the relevance of the test task to the target language domain. Depending on the prior knowledge and experience of the candidate, some test tasks may be more contextualized than others. A candidate's ability to respond may be further enhanced by his ability to activate this relevant information. In a CR test, a domain is specified upon which the score can be validated. The context in that domain may well vary from student to student. While the students agreed (75%) that "the test tasks were appropriate given my ability in English" (survey 2, item 14), problems can develop with validity when the subject matter is more familiar to one candidate than another. When this occurs, there is a tendency to develop more general topics that do not engage the candidate. To avoid this problem when testing, an examiner would need a variety of specific topics available so the candidate can be engaged while avoiding the problem of contextual bias. When examining a candidate it would seem wise to find out the candidate's background and select a topic that not only engages the candidate but avoids the problem of test bias.

4.14 TTEST

A TTEST determines significant differences in mean scores for identical items. A value below .05 indicates significant differences in mean scores between the students before and after the CR test. A one-tailed TTEST was conducted because the results were expected to confirm what the researcher expected. The following seven items were compared:

Survey Item	**Mean before**	**Mean after**
Group testing was possible	2.67	3.00
Students were compared with each other	2.92	2.67
Current English tests forbid students to interact.	2.51	2.25
Tests teach students why they fail	2.74	2.58
The test was a relative competition (You win, I lose)	3.03	2.72
English test results reflect socio-economic status.	2.31	2.53
Test items reflected real-world situations	2.54	3.08

p = 0.212638064 means there is no significant difference between the two groups

4.15

Open Ended Questions

Three open-ended questions were used on both surveys but the response to them was very poor. The following questions were asked:

Was there anything in this survey that was difficult to understand or confusing?

Was there anything about this questionnaire that you thought was problematic?

Is there anything you feel should have been included in the survey?

How would you make English language testing better?

The lack of response to the first two questions would seem to indicate that there was no problem understanding the survey questions. But their lack of response to the third and fourth questions which asks for their input on what else could have been included on the survey could be interpreted two ways. Either they truly didn't have anything to add due to the novelty of doing an educational survey which may be viewed as a cultural anomaly or they took a cavalier approach when asked to think for themselves.

4.16 Summary of Key Findings

1) The students suggest that they are becoming active as opposed to passive learners (survey 2, item 20). However, this may be because the CSAT is over and they are preparing to improve their communication skills to get a job.

2) They suggest that they are developing a positive attitude towards language learning (survey 2, item 4) and that they should be tested on what they know to boost motivation (survey 2, item 19).

3) They perceive that the competitive nature of testing is subsiding after a CR test and that that they are being evaluated on their ability to perform a task (survey 2, item 7).

4) Socio-economic status was an issue with the students as the 6% increase in agreement after a CR test would seem to indicate. However, once CR tests were implemented, students seem to realize that they are being tested for mastery of a criterion and not

against each other. This percentage difference could fall once they become more familiar with CR tests.

5).The students perceive a CR test as being a reliable indicator of their ability (survey 2, item 11, 13) while at the same time, questioning the reliability of NR tests (survey 1, item 3).

6) They perceive group testing as being possible as they were able to interact with a partner (survey1, item 6, survey 2, item 21). This could be beneficial in Korea where classroom size is large.

7) Score interpretation seems more relevant to the students as the test was an interactive activity using real world situations. (survey 2, item 15).

8) They perceived the test tasks as being appropriate for their abilities (survey 2, item14). This implies that they are better able to determine their ability and content mastery. This enhances content validity, something they questioned in NR tests (survey 1, item 3).

(9) They perceive the use of real world tasks as being more relevant which translates into a better performance by the candidates (survey 2, item 10).

5. Conclusion and Recommendations

The opening pages of this project dealt with the current disconnect between CLT methodology and assessment due to a high stakes university entrance exam (CSAT) that tests only reading and listening skills. The project has introduced an alternative testing method (CR test) and focused on the students perceptions of this test. The research results obtained from the students seem to indicate that CR tests are not only feasible but also create positive backwash in the process as the students seem to be moving from passive to active learners. This perception requires more research particularly in a hierarchical society characterized by status. This testing method may also help improve Korean scores in English language proficiency tests, something that is not occurring at the present time.

A paired CR test would allow students to cooperate instead of compete with against one another. By allowing the students too interact in real world situations, authenticity is promoted. More students are better able to interpret their score as they are allowed to do something they are familiar with while having the intimidation factor of an interviewer removed. Such a test would also be practical given the large class sizes in

Korea. Because the project was limited to students, neither the teachers, KICE, nor the public was surveyed on this testing method. Nor are they likely to be familiar with it. It would be appropriate at this time to introduce a CR test into many of the teacher training workshops that are occurring in Korea and determine the teachers' perceptions of the test. By introducing descriptive bands to teachers during training workshops they will learn how to make subjective judgments reliable at the classroom level, and eventually, the CSAT. This may also indirectly move the teachers away from teaching to a test towards teaching to a curriculum as they will now have proper assessment abilities. KICE seems preoccupied with developing a replacement test for the English portion of the CSAT based on Japan's Eiken test (Kang, 2009). While the test change is being applauded on many fronts, its reliability is being questioned as no one is receiving training in how to subjectively grade the test to ensure reliability.

The private education bill in Korea is a staggering $11 billion annually (Kang, 2009). The students agreed that English test results reflect socio-economic status. At the same time they perceived that they were not being tested against one another but only on mastery of a criterion. If the students prepared together for a CR test, instead of studying at a language institute, their scores might be as good or better than if they continued at private institutes as many of these institutes can't teach speaking and writing on a professional level. This area of formal versus informal learning and subsequent self-assessment requires further research but could be a significant benefit for all students in a status conscious country like Korea. While cram schools will always play a significant role in the preparation for the CSAT, a change in language testing methods could save Korean parents thousands of dollars in education costs.

This project has shown the difference between an NR test and a CR test and the positive backwash that that the latter testing method creates. This has been done not only to improve student motivation and test scores but to deal with the constant cycle of testing that occurs in Korea and seems to have very little benefit. It is a custom in Korea to be constantly testing as Koreans seem enamored with numerical rank and progress, a cultural trait that relies heavily on NR tests that force students to compete against each so they can be better segregated. By adopting an alternative testing method, this endless cycle of testing can finally abate as students realize that they can measure their own strengths and progress in English. Hopefully, this student perception may lead towards self-assessment and reduce this demoralizing and unnecessary cycle of testing. Finally this project has tried to introduce a testing method so students not only improve their language proficiency but also learn to cooperate instead of compete with one another.

6. References

Ahn, S. H. (2003). *A case study of a Korean learner*. Retrieved June 27, 2008 from http://asian-efl-journal.com/Dec 2003

Amrein, A. L. and Berliner, D. C. 2002.High Stakes Testing, Uncertainty, and Student Learning.Educational Policy Analysis Archives, 10, 18.

Atkinson, T and Davies, G. (2000). *Computer Aided Assessment and Language Learning*. ICLT4LT.

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

Bachman, L. & Palmer, A.(1996). *Language testing in practice: Designing and developing useful tests* (pp. 17-42). Oxford: Oxford University Press.

Bae, J. S. (2009, April 16). *Schools Show Wide Performance Gap.* Korea Times, p.3

Borg,W. R., & Gall, M. D. (1983). Steps in conducting a questionnaire survey. In *Educational research: An introduction,* (4th ed., pp. 415-435). New York: Longman.

Breen, M. (2007, January 24). Are we living in heaven or hell? Korea Herald, p.13.

Breen, M. (2008, November 14). *Exam Hell is Over*. Korea Times, p.6.

Breen, M. (2004). *The Koreans.* New York: St. Martin's Press.

Brown, H. D. (2004). *Language assessment: Principles and classroom practices* (pp.19-40). New York: Pearson Education.

Brown, J. D. (2001). *Using Surveys in Language Programs*. Cambridge: Cambridge University Press.

Brown, J. D.& Hudson, Th. 2002: Criterion–referenced Language Testing. Cambridge: Cambridge University Press.

Cambridge ESOL. Retrieved July 12, 2009 from http://www.cambridgeesol.org/what-we-do/who/cambridge-first.html

Carr,W. & Kemmis,S. (1986). Becoming critical: education, knowledge and action research. Deakin University Press, Geelong, Vic.

CEO's press for English education reform. November 18, 2006. *Korea Times*, p. 4.

Choi, Seong Hee. (1999). *Teaching English as a Foreign Language in Korean Middle Schools: Exploration of Communicative Language Teaching through Teacher's beliefs* *and Self-Reported Classroom Teaching Practices*. Diss, Ohio State University. Ann Arbor: UMI . AAT 3031198.

Choi, Yun Ha. (2009, April 4). *For Success of Admissions Reform*. Korea Times. p.7.Collecting and Analyzing Qualitative Data .Retrieved Jan.6, 2007 from http;//www.health.auckland.ac.nz/hrmas/resources/html

Cunningham, C. R. (2002). The TOEIC test and communicative competence: Do test score gains correlate with increased competence? Retrieved January 8, 2007 from http://www.cels.bham.ac.uk/resources/essays/Cunndiss.pdf.

Educational Testing Service. (2004). *TOEIC Technical Manual.* Princeton NJ: Educational Testing Service.

Educational Testing Service. (2006B) TOEIC Can–Do Guide. Princeton NJ: Educational Testing Service.

English tests: Comparing the TOEFL and the TOEIC. Retrieved January 16, 2007 from http://www.voanews.com/specialenglish/

Ennis, R. H. (1999). *Test Reliability: A Practical Exemplification of ordinary Language Philosophy*. Philosohy of Education.

ETS. TOEFL. Retrieved January 6, 2007 from http://ets.org/portal/site/ets/menuitem.fab

Flattery, B. (2007). *Language, Culture, and Pedagogy: An overview of English in South Korea.* Retrieved Feb. 4, 2008 from http://www.chass.utoronto.ca/cpercy/ courses/eng6365-flattery.htm

Finch, A. & Shin, D. (2005). *Integrating teaching and assessment in the EFL classroom. A practical guide for teachers in Korea.* Seoul: Sahoipyungnon Publishing Co., Inc.

Gardner, R. C.(1985). *Social Psychology and Second LanguageLearning*. London: Edward Arnold

Hartman, P. (2009, April 16). *Don't let the numbers fool you.* Korea Times p. 8

Heaton, J. B. (1998). Approaches to language testing. *Writing English language tests*. (pp.15-24). London: Longman

Henning, G. (1987). *A guide to language testing: Developments, evaluation, research* (pp.75-80). Boston, Mass: Heinle & Heinle.

Horwitz, E. K. (1985). 'Surveying student beliefs about language learning and teaching in the foreign language methods course.' *Foreign Language Annals*, 18 (4), 333-340.

Hughes, A. (2003). *Testing for Language Teachers.* Cambridge: Cambridge University Press.

IELTS. *Teaching Resources*. Retrieved July 10, 2009 from http://www.cambridgeesol.org/teach/ielts/index.htm

Kang, S. W. (2009, April 14). *Educators Flunk Test Management*. Korea Times, p.3.

Kang ,S. W. (2008, June 3). *Korean English fever betrayed by test scores*. Korea Times. Retrieved July 3,2008 from http:// www.koreatimes.co.kr/news/2008/06/117 html

Kang, S. W. (2008, December 15). *New English Test to Debut for College Admission*. Korea Times, p.1.

Kang, S. W. (2008, December 19). *Korean Version of the TOEFL to Debut*. Korea Times, p.1.

Kang, S. W. (2009, April, 11-12) *State- Run Test Triggers Louder Noise*. Korea Times, p.3.

Kang, S. W. (2009, April 23). *Should Admission Tests Results Be Disclosed*? Korea Times, p.5.

Kemmis, S. & McTaggart R. (eds) 1988, *The action research planner* (pp.5-27). Geelong: Deakin University Press.

Kim, Hyun Sook, (2003). *The types of speaking assessment tasks used by Korean Junior secondary school English teahers.* Retrieved July 29, 2009 from www.asian-efl-journal.com/dec_03_gl.kr.php

Kim, S. J. (2004). *Coping with cultural obstacles to Speaking English in the Korean Secondary School Context.* Retrieved Feb.3, 2008 from http:// asian-elf-journal.com/September_04_ksj.php

Knapman, G. S. (2007). *The TOEIC- a critical review.* Retrieved July 26, 2009 from http://crf.flib.u-fukui.ac.jp/dspace/bitstream/10461/2907/1/pdf

Kohonen, V. (1999). Authentic assessment in affective foreign language testing. In J. Arnold (Ed.), *Affect in language learning*. (pp.279-294). Cambridge: Cambridge University Press.

Kramsch, C. *Third Culture and Language Education*. Chapter 11. Retrieved June 6, 2009 from lrc.cornell.edu/events/past/2008-2009/papers08/third.pdf

Lee, Su Hyun (2007, May 16). *South Koreans jostle to take an English test*. New York Times. Retrieved July 3, 2008 from http:// www.nytimes.com/2007/05/17/world/07//asia

Liddicoat, A. 1999. The Challenge of Intercultural Language Teaching. *Intercultural Literacy - the Key to Cross-Cultural Success.* Retrieved June 3, 2009 from http://www2.asialinkuniblb.edu.au/aef/shanghai/papers/ws4.html

Lynch, R. (2003). *Authentic, performance based assessment in ESL/EFL reading instruction.* Retrieved July 24, 2008 from http://www. asian-elf-journal.com/dec_03

McNamara, T. 2000: *Language Testing*. Oxford: Oxford University Press

Messick, S.(1988). The once and future issue of validity: assessing the meaning and consequences of measurement. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp.33-46). Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Mungubhai, F.(2006). *Language testing study book*. Toowoomba: University of Southern Queensland Press.

Noffke, S. & Somekh, B., (2005). *Action Research*. In B. Somekh & C. Lewin (Eds.). Research methods in social sciences (pp.99-101). London: Sage Publications

Nunan, D. (2005) *Important Tasks of English Education: Asia-wide and Beyond.* Retrieved from http://www. asian-elf-journal.com/September_05_dn.php.

Nunan, D. (1999). Second language teaching and learning. Heinle and Heinle publishers.
Nunn, R. (2007). *Redefining Communicative Competence for International and local Communities*. Retrieved June 27, 2008 from http://asian-efl-journal.com/Dec 2007

Pajares, F.(2007). *The elements of a proposal.* Retrieved June, 3, 2008 from http://des. emory.edu/mfp/proposal.html

Park, S. (2006). *The Impact of English Language and Cultural Variations on Korean Students in Australian Undergraduate Programs*. Retrieved June 2, 2009 from http://eprints.usq.edu.au/view/people/Park=3ASang

Pin, Yi Ching. *A critical review of five language washback studies from 1995-2007: methodogical considerations*. Retrieved July 22, 2009 from http:// jalt.org/test/plan_1.htm

Prapphal, Kanchana (2008). *Issues and trends in language testing and assessment in Thailand.* Retrieved March 24, 2008 from http://www.ciillibrary.org:8000/ciil/full text/ Language_testing_Vol_24_no_1

Richards, J. C. (2006). *Curriculum development in Language Teaching.* Cambridge: Cambridge University Press

Richards, J. C. & Rogers, T. S. (2005). *Approaches and Methods in Language Teaching*. Cambridge: Cambridge University Press.

Ruffin, Rick. (2009, April, 5). *Dubious Distinction*. Korea Times p.7.

Saito, Y. (2009). *The Use of Self-Assessment in Second Language Assessment.* Retrieved April 4' 2009 from http://www.tc.columbia.edu/academic/tesol/WJfiles/pdf/Saito

Sewell, H. D. (2005). *The TOEIC: Reliability and Validity within the Korean Context*. Unpublished.

Stevens, R. (2009, May 5). *Why Few Korean Master English*. Korea Times, p.9.

Taylor, L. (1999): *Study of Quantitative Differences between CPE Individual and Paired Test Formats,* Internal UCLES EFL Report.

Tuckman, B.W. (1994). *Conducting educational research*, 4th edn., Orlando: Harcourt Brace & Co.

USQ Study book (2006). *Research methods in education.* Toowoomba: University of Southern Queensland Press.

Web center for Social Research methods. Retrieved January 12, 2007 from http://www.socialresearchmethods.net/kb/qualdeb.php

7. APPENDIX

7.1 Student Survey 1

설문조사 1

Introduction

The purpose of this survey is to get your opinion about English language testing in
이 조사의 목적은 한국에서 영어능력테스트에 대한 여러분의 의견을 얻기 위함
Korea. Your opinion about English language testing will be used for private
입니다. 설문조사에 대한 여러분의 의견은 개인적인 업무에 사용되어질 것입니다.
research. Your name will remain anonymous. Thank you for participating in this survey.
여러분의 이름은 익명으로 남을 것이고 이번 조사에 참여해 주신 것을 감사하게 생각합니다.

Date survey was taken(mm/dd/yyyy):__/__/____Age:
날짜(월/일/년도): __/__/____나이:

Gender (circle one): M F
성별(남,여)

Please circle an appropriate response.
다음을 읽고 적당한 답에 동그라미 치세요.

1.When I take a test in English, I focus on only one correct answer.
1.나는 영어시험을 칠 때, 오직 하나의 정답에만 중점을 둔다.

Strongly disagree	Disagree	Agree	Strongly Agree
매우 그렇지 않다	**그렇지 않다**	**그렇다**	**매우 그렇다**

2.Time limits do not allow me time to finish my test
2.시간 제한 때문에 나는 시험을 다 끝내지 못한다.

Strongly disagree	Disagree	Agree	Strongly Agree
매우 그렇지 않다	**그렇지 않다**	**그렇다**	**매우 그렇다**

3. Multiple choice exams have similar items to those I studied in class.
3.객관식 유형의 시험은 내가 수업 시간에 공부했던 것과 유사하다.

Strongly disagree	Disagree	Agree	Strongly Agree
매우 그렇지 않다	**그렇지 않다**	**그렇다**	**매우 그렇다**

4. I do not have to speak or write in an English test.
4.나는 영어시험에서 말을 하거나 글을 쓸 필요가 없다.

Strongly disagree	Disagree	Agree	Strongly Agree
매우 그렇지 않다	**그렇지 않다**	**그렇다**	**매우 그렇다**

5. Tests emphasize more what students cannot do than what they can do.
5.영어시험은 학생들이 할 수 있는 것보다 할 수 없는 것을 더 강조한다.

Strongly disagree	Disagree	Agree	Strongly Agree

매우 그렇지 않다 그렇지 않다 그렇다 매우 그렇다

6.Group testing is possible.

6.그룹별 시험이 가능하다.

Strongly disagree Disagree Agree Strongly Agree

매우 그렇지 않다 그렇지 않다 그렇다 매우 그렇다

7.Multiple choice exams test only lower order knowledge.

7.객관식 유형의 시험은 단순지식만 평가한다.

Strongly disagree Disagree Agree Strongly Agree

매우 그렇지 않다 그렇지 않다 그렇다 매우 그렇다

8. English test results reflect socio-economic status.

8.시험결과는 사회-경제적 위치를 반영한다.

Strongly disagree Disagree Agree Strongly Agree

매우 그렇지 않다 그렇지 않다 그렇다 매우 그렇다

9.Students should do a practice test before the actual test.

9.학생들은 실제 시험을 치기 전에 연습 시험을 쳐야만 한다.

Strongly disagree Disagree Agree Strongly Agree

매우 그렇지 않다 그렇지 않다 그렇다 매우 그렇다

10.The test offers the student a variety of different items.

10.영어시험문제유형은 다양하게 출제된다.

Strongly disagree Disagree Agree Strongly Agree

매우 그렇지 않다 그렇지 않다 그렇다 매우 그렇다

11. Current English tests are stressful.

11.현재 영어 시험은 내가 긴장되게 만든다.

Strongly disagree Disagree Agree Strongly Agree

매우 그렇지 않다 그렇지 않다 그렇다 매우 그렇다

12.The test was a relative competition. (You win, I lose).

12.영어시험은 경쟁적인 상대평가이다.

Strongly disagree Disagree Agree Strongly Agree

매우 그렇지 않다 그렇지 않다 그렇다 매우 그렇다

13.Current English tests forbid students to interact.

13.현재 영어시험은 학생들이 상호작용하는 것을 불가능하게 한다.

Strongly disagree Disagree Agree Strongly Agree

매우 그렇지 않다 그렇지 않다 그렇다 매우 그렇다

14. The test tasks reflected real-world situations.

14.영어시험문제는 실제 상황을 연출한 것으로 출제한다.

Strongly disagree Disagree Agree Strongly Agree

매우 그렇지 않다 그렇지 않다 그렇다 매우 그렇다

15.Tests teach students why they fail.

15.영어시험을 통해 무엇이 부족한지 알 수 있다.(시험에 실패한 이유를 알게 한다.)

Strongly disagree Disagree Agree Strongly Agree

매우 그렇지 않다 그렇지 않다 그렇다 매우 그렇다

16. It is embarrassing to speak in English.

16.영어로 말하는 것이 부끄럽거나 어색하다.

Strongly disagree Disagree Agree Strongly Agree

매우 그렇지 않다 그렇지 않다 그렇다 매우 그렇다

17. Students are compared with each other.

17.학생들은 서로 비교가 된다.

Strongly disagree Disagree Agree Strongly Agree

매우 그렇지 않다 그렇지 않다 그렇다 매우 그렇다

18. I worry about making mistakes.

18.나는 실수하는 것에 대해 걱정을 한다.

Strongly disagree Disagree Agree Strongly Agree

매우 그렇지 않다 그렇지 않다 그렇다 매우 그렇다

19. Teachers can test exactly whether students can communicate with each other.

19.교사는 정확하게 의사소통을 할 수 있는지 테스트 할 수 있다.

Strongly disagree Disagree Agree Strongly Agree

매우 그렇지 않다 그렇지 않다 그렇다 매우 그렇다

20. Was there anything in this survey that was difficult to understand or confusing?

20.설문조사 내용 중에 이해하기 힘든 내용이나 헷갈리는 내용이 있었습니까?

____

21. Was there anything about this questionnaire that you thought was problematic?

21.설문조사 문항 중에 의문점이나 문제가 될 만한 것이 있었습니까?

____

22. Is there anything you feel should have been included in the survey?

22.설문조사 문항 중에 추가하고 싶은 부분이 있습니까?

____

7.2 Criterion Referenced Test. Survey 2

설문조사 2

The purpose of this survey is to get your opinion about English language testing in

이 조사의 목적은 한국에서 영어능력테스트에 대한 여러분의 의견을 얻기 위함

Korea. Your opinion about English language testing will be used for private

입니다. 설문조사에 대한 여러분의 의견은 개인적인 업무에 사용되어질 것입니다.

research. Your name will remain anonymous. Thank you for participating in this survey.

여러분의 이름은 익명으로 남을 것이고 이번 조사에 참여해 주신 것을 감사하게 생각합니다.

Date survey was taken(mm/dd/yyyy):__/__/____

날짜(월/일/년도): __/__/____

Age: Gender (circle one): M F

나이: 성별(남,여)

Years of language teaching experience in a formal context:____ years

학교나 학원에서 영어를 배운 기간: ____년

Please circle an appropriate response.

다음을 읽고 적당한 답에 동그라미 치세요.

1. The test allowed me to be involved in a variety of tasks.

1.이번 시험은 내가 영어의 말하기,듣기,읽기 등 다양한 부분을 공부할 수 있게 하였다.

Strongly disagree Disagree Agree Strongly Agree

매우 그렇지 않다 그렇지 않다 그렇다 매우 그렇다

2. **The test tasks reflected real world situations**

2.테스트 항목들은 실생활을 반영하였다.

Strongly disagree Disagree Agree Strongly Agree

매우 그렇지 않다 그렇지 않다 그렇다 매우 그렇다

3. My test score reflected my abilities in English

3.나의 시험 점수는 실제 내 영어 능력을 반영했다.

Strongly disagree Disagree Agree Strongly Agree

매우 그렇지 않다 그렇지 않다 그렇다 매우 그렇다

4. The test motivated me to perform the task required.

4.그 시험은 나에게 주어진 업무를 수행할 수 있는 동기를 부여하였다.

Strongly disagree Disagree Agree Strongly Agree

매우 그렇지 않다 그렇지 않다 그렇다 매우 그렇다

5. There was more than way to answer the question.

5.시험문제에 답 할 수 있는 더 많은 방법들이 있었다.

Strongly disagree Disagree Agree Strongly Agree

매우 그렇지 않다 그렇지 않다 그렇다 매우 그렇다

6. One word answers were not appropriate.

6.단답형 답은 적절하지 않았다.

Strongly disagree Disagree Agree Strongly Agree

매우 그렇지 않다 그렇지 않다 그렇다 매우 그렇다

7. The test was a relative competition (You win, I lose).

7.이번 시험은 상대적 평가였다.

Strongly disagree Disagree Agree Strongly Agree

매우 그렇지 않다 그렇지 않다 그렇다 매우 그렇다

8. I was evaluated on my ability to perform a task.

8.시험은 학습능력을 평가했다.

Strongly disagree Disagree Agree Strongly Agree

매우 그렇지 않다 그렇지 않다 그렇다 매우 그렇다

9. Students were compared with each other.

9.학생들은 서로 비교되었다.

Strongly disagree Disagree Agree Strongly Agree

매우 그렇지 않다 그렇지 않다 그렇다 매우 그렇다

10. The test measured my progress in English

10.이 시험은 나의 영어실력을 평가하였다.

Strongly disagree Disagree Agree Strongly Agree

매우 그렇지 않다 그렇지 않다 그렇다 매우 그렇다

11. I was able to determine what I can and cannot do in English

11.이번 시험을 통해 나의 영어실력의 정도를 알 수 있었다.

Strongly disagree Disagree Agree Strongly Agree

매우 그렇지 않다 그렇지 않다 그렇다 매우 그렇다

12. The in-class tasks and the test tasks were similar.

12.수업시간에 배우고 있는 영어수준과 시험수준이 비슷하였다.

Strongly disagree Disagree Agree Strongly Agree

매우 그렇지 않다 그렇지 않다 그렇다 매우 그렇다

13. The test tasks set standards of learning that were achievable

13.영어시험이 영어실력을 키울 수 있는 기준을 만들었다.

Strongly disagree Disagree Agree Strongly Agree

매우 그렇지 않다 그렇지 않다 그렇다 매우 그렇다

14. The test tasks were appropriate given my ability in English.

14.이번 시험이 내 영어실력에 적합하였다.

Strongly disagree Disagree Agree Strongly Agree

매우 그렇지 않다 그렇지 않다 그렇다 매우 그렇다

15. The test emphasized my strengths and progress in English

15.영어 시험을 통해 말하기,듣기,쓰기,읽기 중 취약점을 알 수 있었다.

Strongly disagree Disagree Agree Strongly Agree

매우 그렇지 않다 그렇지 않다 그렇다 매우 그렇다

16. My confidence was enhanced by taking this test.

16.이번 시험을 통해 영어에 대한 자신감이 조금은 올라갔다.

Strongly disagree Disagree Agree Strongly Agree

매우 그렇지 않다 그렇지 않다 그렇다 매우 그렇다

17. I was allowed to be successful

17.이번 시험을 잘 본 것 같다.

Strongly disagree Disagree Agree Strongly Agree

매우 그렇지 않다 그렇지 않다 그렇다 매우 그렇다

18. English test results reflect socio-economic status.

18.이번 시험은 학과수업 외에 따로 영어를 배우는 학생들에게 유리한 것 같다.

Strongly disagree Disagree Agree Strongly Agree

매우 그렇지 않다 그렇지 않다 그렇다 매우 그렇다

19. The test allowed me to think and respond to a question.

19.이번 시험에서 질문에 대해서 생각하고 답을 할 수 있었다.

Strongly disagree Disagree Agree Strongly Agree

매우 그렇지 않다 그렇지 않다 그렇다 매우 그렇다

20. Active awareness of learning was promoted

20.영어를 좀 더 적극적으로 배우고 싶은 마음이 생겼다.

Strongly disagree Disagree Agree Strongly Agree

매우 그렇지 않다 그렇지 않다 그렇다 매우 그렇다

21. Group testing was possible

21.그룹테스트가 가능하였다고 생각한다.

Strongly disagree Disagree Agree Strongly Agree

매우 그렇지 않다 그렇지 않다 그렇다 매우 그렇다

22. Fluency was more important than accuracy

22.말하기 시험에서는 문법의 정확성을 따지기보다 자신감을 가지고 말하는 것이 더 중요한 것 같다.

Strongly disagree Disagree Agree Strongly Agree

매우 그렇지 않다 그렇지 않다 그렇다 매우 그렇다

23. Current English tests forbid students to interact.

23.현재 영어시험은 학생들이 상호작용하는 것을 불가능하게 한다.

Strongly disagree Disagree Agree Strongly Agree

매우 그렇지 않다   그렇지 않다   그렇다   매우 그렇다

24. Tests teach students why the fail..

24.시험은 학생들에게 성적이 오르지 않는 이유를 설명해준다.

Strongly disagree   Disagree   Agree   Strongly Agree

매우 그렇지 않다   그렇지 않다   그렇다   매우 그렇다

25. Was there anything in this survey that was difficult to understand or confusing?

25.이번 조사에서 이해하기 어렵거나 혼란스러운 질문이 있었습니까?

26. Was there anything about this questionnaire that you thought was problematic?

26.이번 조사에서 결정하기 어려운 질문은 몇 번입니까?

27. Is there anything you feel should have been included in the survey?

27.이번 조사에서 꼭 포함되었으면 하는 질문은 무엇입니까?

28. How would you make English language testing better?

28.당신은 앞으로 영어실력 향상을 위해 어떻게 할 것입니까?

7.3 Statistical Table of Results

Table 1. Student Survey 1.

Survey Item	Mean	% agree	Std. Dev.
*Item 1***.** When I take a test in English, I focus on only one correct answer.	2.44	61	.821
Item 2. Time limits do not allow me to finish my test.	2.10	53	.641
Item 3. Multiple choice exams have similar items to those I studied in class.	2.13	53	.362
Item 4. I do not have to speak or write in an English test.	2.13	53	.615
Item 5. Tests emphasize more what students cannot do than what they can.	2.10	53	.502
Item 6. Group testing is possible.	2.67	68	.577
Item 7. Multiple choice exams test only lower order knowledge.	2.33	58	.577
Item 8. English test results reflect socio-economic status.	2.31	58	.731
Item 9. Students should do a practice test before the actual test	2.69	67	.614
Item 10. The test offers a variety of different items.	2.72	68	.510
Item 11. Current English tests are stressful.	3.03	76	.486
Item 12. The test was a relative competition (you win, I lose).	3.03	76	.510
Item 13. Current English tests forbid students to interact.	2.51	63	.556
Item 14. Test items reflected real-world situations.	2.52	63	.643
Item 15. Tests teach students why they fail.	2.74	69	.549
Item 16. It is embarrassing to speak in English.	2.74	69	.677
Item 17. Students were compared with each other.	2.92	73	.532
Item 18. I worry about making mistakes.	2.95	74	.686
Item 19. Teachers can test exactly whether students can communicate with each other.	2.68	67	.667

7.4 Statistical Table of Results

Table 2. Criterion Referenced Test Student Survey.

Survey Item	Mean	% Agree	Std. Dev.
*Item 1***.** The test allowed me to be involved in a variety of tasks.	3.08	77	.439
Item 2. Test items reflected real world situations.	3.08	77	.500
Item 3. My test score reflected my abilities in English.	2.94	74	.538
Item 4. The test motivated me to perform the task required.	3.03	76	.506
Item 5. There was more than one way to answer a question.	3.06	77	.583
Item 6. One word answers were not appropriate.	2.86	72	.762
Item 7. The test was a relative competition (you win, I lose)	2.72	68	.506
Item 8. I was evaluated on my ability to perform a task	2.61	65	.506
Item 9. Students were compared with each other.	2.67	67	.717
Item 10. The test measured my progress in English.	3.03	76	.560
Item 11. I was able to determine what I can and cannot do in English.	3.11	78	.575
Item 12. The in-class and test tasks were similar.	3.08	77	.554
Item 13. The test tasks set standards of learning that were achievable	2.94	74	.630
Item 14. The test tasks were appropriate given my ability in English.	2.99	75	.465
Item 15. The test emphasized my strengths and progress in English.	3.14	79	.593
Item 16. My confidence was enhanced by taking this test.	3.03	76	.609
Item 17. I was allowed to be successful.	2.53	63	.878
Item 18. English test results reflect socio-economic status.	2.53	63	.810
Item 19. The test allowed me to think and respond to a question.	3.08	77	.500
Item 20. Active awareness of learning was promoted.	3.22	81	.591
Item 21. Group testing was possible.	3.00	75	.586
Item 22. Fluency was more important than accuracy.	3.61	90	.599
Item 23. Current English tests forbid students to interact.	2.25	56	.792
Item 24. Tests teach students why they fail	2.58	65	.768