Modern English Education, Vol. 12, No. 1, Spring 2011

Validation of Level Classification in an English Proficiency Test for Young-Learners*

Haewon Pyo, Haedong Kim†
Hankuk University of Foreign Studies

Pyo, Haewon & Kim, Haedong. (2011). Validation of level classification in an English proficiency test for young-learners. *Modern English Education*, *12*(1), 73-88.

This study aims to illustrate how to determine and justify level classifications for a standardized test for young learners. The data from eight test developers and 1,378 test takers were utilized for the analysis. The Angoff method was used to get a pass/fail boundary and the identified cut score was compared with the actual scores obtained from the test. Correlation coefficients were respectively checked on listening questions and reading questions. Cluster analyses were carried out to identify the number of possible levels, which were checked by the test writers' predicted scores on each item. The results indicate that there were two distinguishable level-groups. However, there was a gap between expected cut score and actual cut score. As a conclusion, it is proposed to use the methods used in this study to check the validity of level classification in a standardized test.

[testing/level/cut score/validation /평가/수준/분리 점수/타당화]

I. INTRODUCTION

This study aims to illustrate the methods and procedures for setting up and justifying a pass/fail boundary, i.e. a cut score, for a standardized test for young learners. In the context of testing, establishing a cut score is often regarded as an important process in designing a test (Weir, 2005). Downing (2006) says that the methods and procedures used to establish a cut score are a major source of validity evidence. The pass/fail boundary represented by a

* This study includes parts of the first author's unpublished MA thesis and the corresponding author's presentation at KATE 2008 conference. This work was supported by Hankuk University of Foreign Studies Research Fund of 2010.

† Haewon Pyo: first author; Haedong Kim : corresponding author.

cut score means a minimally accepted ability of test takers (Cizek, 2006). By setting up the boundary of a cut score under a systemic and reliable process, test takers' ability can be reliably identified and this will enhance the validity of score interpretation.

Before setting up a cut score, information on the test takers' language competence should be obtained. And after the administration of the test, the results of the test takers' competence should be analyzed to check the accuracy of the cut score. However, few studies have reported the comparison between an expected cut score and a real cut score. This justifies the value of the present study.

Like assessment targeting adolescents or adults, assessment for young learners should be cautiously developed and implemented due to their specific characteristics. McKay (2006) said that young learners are under the process of active development physically as well as mentally and that they are very vulnerable to the radical changes in their external environment. However, in Korea, even though over a decade has passed since English education was established in elementary schools, it can be noted that there has been a lack of practical and empirical research on testing tools. In order to determine if the assessment measuring Korean young learners' English language proficiency is valid, then it would be necessary to conduct research supporting the validity argument of a test. This also supports the need of the present study.

The test used in the study was mainly developed and implemented as a form of a contest. The ultimate goal of the test is to increase the level of English language education in Korea by providing explicit information about Korean young learner's current English proficiency.

II. LITERATURE REVIEW

1. Validity

Validity assessment is an essential requirement in the field of test development. Validity in assessment reflects the need to evaluate the test and to see whether the test measures what it was designed to measure. To evaluate a test it must be given to the target population and the scores obtained need to be interpreted. In other words, the validation process is necessary in proving the quality of a test. Hughes (2003) suggested that it is required for any test to provide validity evidence. Even though a lot of test makers accept this premise, there have been few tests that showed satisfactory evidence of validity. Messick (1989) insisted that the definition of validity is "the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of interpretations and actions based in test scores" (p. 13). However, interpretation of the degree by which a test should be evaluated is different among researchers and experts;

consequently, as McNamara (1996) noted, it is like 'opening the Pandora's box.' Validation is not a simple judgment of "yes" or "no", but it is an argument in process which needs constant verification; all sources of evidence must complement each other.

In the literature, there are several researchers who propose various ways of validation. Weir (2005) provided a two-sided perspective such as a priori evidence gathered before the administration of a test event and a posteriori evidence gathered after the administration of the test. This process emphasizes the view that validity evaluation is essential in the test development process. He included theory-based validity and context validity in a priori validity evidence and scoring validity. Downing (2006) also proposed a way to provide validity evidence using the twelve steps of test development. According to him, in order for a test to achieve a high validity score, the implementation of test development should be based on relevant theoretical and educational principles of assessment. The twelve steps provide a convenient organizational framework for collecting and reporting all sources of validity evidence for a testing program. They elaborate detailed tasks that should be accomplished in each level and produce validity evidence to be documented in a report that describes the whole development process as well as the result analysis. Others (e.g.. Bachman & Palmer, 2000; Chapelle, Jamieson, & Hegelheimer, 2003; Keyser & Sweetland, 1984-1994) also suggest processes to determine test validation.

Within the Korean context, studies on principles and techniques for test validation have been relatively neglected. Dong-Il Shin (2002) developed five level descriptors for speaking and writing abilities of high school students based on analysing 50 students' performance. But this study did not focus on young learners. In the case of Dong-Il Shin, Ji-Hyun Song and Na-Hee Kim (2006), they explored standardized tests in Korea to compare how the assessments take account of the special characteristics of young learners. They, however, did not use quantitative analysis. They used content analysis on item types, test criteria and level descriptors. It would be useful to propose and to conduct empirical analysis checking the validation of level descriptions and/or classifications.

The present study focuses on test items. We will adapt Weir's (2005) idea of checking the relationship between a priori evidence gathered before administration of a test event and a posteriori evidence gathered after the administration of the test. Evaluating validity of a test through the analysis of test items is an important factor to be considered. Tae-Jae Seong (2005) also said that validity and reliability of tests are supported by the analysis of their items. That is, to evaluate the validity of tests, the analysis of their items should be implemented.

2. Tests for Young Learners

1) Positive and Negative Effects of Tests

Large-scale tests have been developed to target numerous learners across school districts or school systems. Implementing large-scale tests to young learners, however, is very controversial. Opponents insist that it is hard to get prompt feedback that both teachers and learners can use (McKay, 2006). Young learners cannot recognize the test process and the importance of the results easily. This might decrease the reliability of a test's results (Hasselgreen, 2000). Moreover, it has been argued that there is a possibility that young learners' education would be distorted since schools and teachers who worry about the results may focus more on improving test wiseness rather than real knowledge.

On the other hand, there are stalwart supporters (e.g. Gottlieb, 2003) who insist that tests have a positive influence. Their main idea is that if a test has a sound framework and valid content, then they will play a crucial role in education in conjunction with a mandated or published curriculum. By utilizing data elicited from the tests, educational resources can be effectively allocated as well. This might help to set up a more improved educational foundation.

2) Reality of Tests for Young Learners

In the U.K., there are two main organizations, the University of Cambridge Local Examinations Syndicate (UCLES) and the London Chamber of Commerce and Industry Examinations Board (LCCIEB), which provide data related to English language assessment. Based on their research, several English language tests such as the Key English Test (KET), Preliminary English Test (PET), First Certificate in English (FCE), and Cambridge Young Learners English Test (CYLET) have been developed.

Among them, CYLET was administered in 1997 for the first time. This test targets children ranging from ages 7 to 12 who are learning English as a foreign language around the world. Its overall purpose is to assess learners' English language abilities and progress under the consideration of sampling relevant and meaningful language use. It also aims to promote effective learning and teaching, to encourage future learning and teaching of English and to measure proficiency accurately and fairly. There are three levels of CYLET: Starters, Movers, and Flyers, and all of them are aligned at levels A1 (Breakthough) and A2 (Waystage) according to the Council of Europe's Common European Framework for Modern Languages. Starters is designed for learners who are 7 years old with over 100 hours of English lessons, Movers for 8 to 11 with over 175 hours of English lessons, and Flyers for 9 to 12 with 250 hours of English lessons. However, even though the test is developed with careful consideration of young learners' features, the limitation is that it does not reflect the Korean context in which Korean test takers learn English. This may cause a decrease in the validity of the test when it is used for Korean test takers only.

Large-scale English language tests for young learners began to be developed in Korea in

2001. Later, because of increased interest in English education in primary schools, numerous tests targeting young learners such as the Practical English Language Test for Junior (PELT Jr.), Language Arts Testing and Training for Junior (LATT Jr.), Test of the Skills in the English Language (TOSEL), Spoken English Proficiency Test for Junior (SEPT Jr.), Junior English Test (JET), General Tests of English Language Proficiency for Junior (G-TELP Jr.), English Speaking Proficiency Test for Junior (ESPT Jr.), and TOEIC Bridge were introduced.

However, there is a limited amount of research which reports their validity and reliability through analysis. So-Young Koo and Hae-Dong Kim (2007) reported the results of item analysis on a LATT Junior Test in aiming to contribute to improvement of the English tests for primary school students. Hee-Jung Jung (2008) investigated the variables that can affect the difficulty of English reading test items within LATT tests for middle school students. Questionnaires were used to compare the students' expected difficulty levels based on the developer's intuition, with that of the actual test results. The variables found to be strongly correlated to the proportion of correct responses were test items related to 'understand titles' and 'understand reference.'

However, these studies did not attempt to check the relationship between expected and actual level classifications. Gathering reliable information related to young learners' language proficiency should be important and establishing valid assessment tools are inevitable. There should be more research on validity argument of tests in order to improve the current assessment situation and to set up an effective language educational framework. Therefore, research on a valid way of establishing cut scores is critical. This aspect justifies the needs for the present study.

3. Cut Score

Since cut score has a big impact on test takers, establishing relevant passing scores, or cut scores, for the test is a critical process in test development. Before one can determine the scores, a lot of information about the test takers' abilities has to be carefully gathered and a strong legitimacy has to be established as well. Downing (2006) said the cut scores are outlined in the early stage of test development and the shape of the scores is roughly organized through various trials and researches that occur during the development process. Since the results which provide data related for every test taker's ability ought to be combined together to elicit the accurate borderline of cut scores, the real cut scores can only be clearly made after the test has been administered.

In the literature, various ways of establishing cut scores have been proposed (Downing, 2006). One of the most commonly used methods is the Angoff method. It requires the expected passing score judgment in each individual test question. Another way of

establishing the borderline of cut scores is the Eble method. The estimation of the score is made according to the information gathered from the classified categories of difficult and relevant items. The Hofstee method adapts the expected maximum and minimum passing and failure scores that are determined by the reflections of experts on the actual test data. Among the methods, the present study employs the Angoff method to determine the predicted cut score. The main reasons that the Angoff method is being used are that it is relatively well-known and that it can be easily implemented. As it is not much involved in sophisticated statistical procedures, it can be well adopted by a group of test writers.

In statistics, to distinguish groups of cases, i.e. test items that have a similar profile of test scores, cluster analyses were conducted. Unlike factor analysis - which usually attempts to confirm or to identify a similar construct of test items - cluster analysis is helpful to identify groups of cases that have a similar profile of scores across a set of variables (Scholfield, 1998). Data clustering is a common technique for statistical data analysis, which is used in many fields, including machine learning, data mining, pattern recognition, image analysis and bioinformatics (Romesburg, 2004). In the field of ELT, not many studies have used cluster analysis (Scholfield, 1998). The present study attempts to use cluster analysis to sort through the expected test scores and to group the test-items into clusters of different score-groups.

III. RESEARCH METHOD

1. Instrument

The main data used in this research were test prompts and responses from an English proficiency test that was implemented on October 14th, 2007. This test was divided into two levels Test-A and Test-B. Test A was made for third and fourth graders and Test B was for fifth and sixth graders. Each test had three sections, listening, reading, and writing. There were 40 items in each listening section, 20 items in each reading section, and 4 items in each writing section.

Among them, items in the listening and reading sections were chosen as research data because they were presented in a multiple choice format. The items in the writing section which required constructed responses were excluded in the present analysis. Items in the data were arranged in linear order for the convenience of analysis. So numbers from 1 to 40 designated listening items and numbers from 41 to 60 designated reading items. Learners had 30 minutes to answer the questions in the listening section and had 40 minutes to answer the questions in the reading section.

For the analysis, data from 548 applicants in Test-A and 830 applicants in Test-B were utilized. All of them were Korean elementary school students, ranging from 10 to 13 years

of age. Test takers' regional backgrounds vary because the test was administered in five different cities: Seoul, Daejeon, Gwangju, Daegu, and Busan.

Level description of the test was mainly developed on the basis of Korean young learners' target language use (TLU) domain. The domain of the test included real life language usage as well as academic language usage. Task types included listening to dialogues of diverse lengths, reading articles on various topics, writing sentences using simple vocabulary and carrying out basic conversations. The ultimate goal of these tasks was to reflect authentic language usage in a real life context. In some part, it also reflected the National Curriculum in Korea because it targets Korean young learners as its main test takers. The Korean education system follows the National Curriculum which provides the fundamental framework like achievement levels. The areas and tasks of achievement level in the National Curriculum affected to the TLU domain of the level description.

Even though the framework of the level descriptions is distinguished by age groups and school years, every test taker can take the appropriate level of the test if he or she wants. In considering this aspect, the framework of the level description is organized based not on ages or school years but on test takers' language proficiency. Table 1 presents the level description of the test for 3/4 graders and 5/6 graders.

Table 1
Level Descriptions of the Test

Levels	Skills		Description
5/6 graders (Test-B)	comprehension	listening	- Understand the main ideas and details from various conversations related to daily life - Recognize the relationships among people who take part in the conversation and grasp the key contexts - Carry on tasks after getting information through medium-length listening texts
		reading	- Understand the main ideas and details of medium-length texts related to daily life - Grasp the flow and logical organization of texts - Manage tasks based on information presented in the tables, charts, graphs and pictures - Infer the meaning of difficult vocabulary or phrases from given texts - Recognize simple cohesion among sentences
	expression	speaking	- Describe familiar people or objects - Produce detailed and maintain basic, coherent conversation related to daily life - Express information about tables, charts, graphs and pictures - Carry on simple tasks through conversations with others - Describe past as well as future events
		writing	- Write simple letters or diaries - Summarize simple texts into a few sentences - Appropriately use punctuation

3/4 graders (Test-A)	comprehension	listening	- Recognize simple explanation of familiar objects - Recognize intentions embedded in simple conversations as well as detailed content - Carry on tasks after listening to simple explanations - Recognize past, present, and future tenses of given information
		reading	- Recognize main subjects and the detailed information of simple texts about daily life -Recognize past, present and future tense of given information
	expression	speaking	- Appropriately interact with others and express opinions through simple conversations - Express main ideas and detailed information after listening to simple explanations
		writing	- Briefly express past, present, and future events - Write letters or keep diaries using simple vocabulary or a few sentences

Not all target-level descriptions of the National Curriculum were covered in the present level descriptions of the test. However, as Young-Shik Lee and Hye-Young Kim (2009) attempted, the test made an attempt to reflect the Korean context in order to enhance the content validity of the test.

2. Participants

There were eight test developers who participated in developing the proficiency test and all of them were assigned different roles in test development process such as level description developer and illustrator, listening item developer, reading item developer, and proof reader. Four item writers developed the reading items and the other four writers developed the listening items. There were two native English speakers who wrote test items and took the role of proof readers.

All of the members are experts in English language education and have at least five years of teaching experience. Half of them are English instructors in universities and two of them are teacher-researcher for young learners in Korea. In general, it was assumed that the test developers were experienced practitioners in the field of English language teaching.

3. Procedures

The test scores obtained from Test A and B were utilized for the analysis. First, an ideal mean test score of the group, proposed by the test developers, was compared with the actual mean score. In other words, in this study, the Angoff method was used to get a cut score and it was compared with the actual score obtained from the test. The Angoff method aims to find out whether the item writers had a relatively accurate understanding of the test

takers' English language ability when they developed the items. If the item writer's estimation of the test takers' abilities was too high or too low, there would be a large gap between the predicted score and the real score. The predicted score of each item, ranging from 0 to 100, was made by the eight item writers. Each individual writer's predicted score was added and divided to elicit the mean score—it was a predicted cut score that was compared to the mean of real scores. Pearson r was chosen to measure correlation coefficient between the predicted score and the actual score.

Since the test item writers cooperated together to write, crosscheck and revise items, it was almost impossible to single out one specific writer for each item. Therefore, the predicted scores for the listening items were gathered from all item writers in the listening part and the same was done for the reading part. As the listening and reading questions were multiple choice items, rater reliability was not checked. Therefore, a multi-faceted Rasch[1] analysis, which is a good way to identify the rater differences among many raters (Weir, 2005) was not carried out. In other words, the reliability check was beyond the scope of the present study.

Second, cluster analyses were conducted to identify the number of possible levels, which were checked by the test writers' predicted scores on each item. The two analyses using Ward method and the K-means method were carried out for cluster analysis, because it is usually recommended to try different methods of cluster analyses to enhance the reliability of the results (Scholfield, 1998). The overall aim of the cluster analysis was to check the suitability of level division.

IV. RESULTS AND DISCUSSION

Table 2 shows the comparison of the predicted and real mean scores in Test-A and Test-B. In the case of Test-A, the mean predicted score in the listening section which consists of 40 items is 21 points higher than the mean real score. In the case of reading section which has 20 items, the predicted mean score is 17.25 points higher than the real mean score. A similar results pattern was observed in case of Test-B.

In case of Test-B, the predicted mean score in the listening section is 82.20 and the standard deviation is 5.10. This is higher than those of the real mean score which were 61.22 and 20.95. It can be noticed that the gap between the two mean scores, 20.98, is similar to the gap in the level 3/4 (Test-A), 21. In the reading section, the predicted mean

[1] We appreciate the idea of using a multi-faceted Rasch for further analysis, which was suggested by Prof. Kyung-Hyun Pyo, at the corresponding author's presentation at KATE 2008 conference.

score is 71.06 and the standard deviation is 10.47, yet the real mean score is much lower, 54.62, which shows 16.44 points of mean score gap similar to 17.25. The standard deviation of the real mean score is 21.51.

Table 2
Comparison of the Predicted and Real Mean Scores

Test-A (3/4 grader)	Mean	SD	N
Predicted listening scores	83.18	4.47	40
Real listening scores	62.18	19.69	40
Predicted reading scores	66.87	14.03	20
Real reading scores	49.62	19.70	20
Test-B (5/6 grader)	Mean	SD	N
Predicted listening scores	82.20	5.10	40
Real listening scores	61.22	20.95	40
Predicted reading scores	71.06	10.47	20
Real reading scores	54.62	21.51	20

Table 3 shows the result of correlation coefficient between the predicted mean score and the real mean score in each test set. The highest level of correlation .579 is shown in level 3/4 whereas the lowest was .491 in the listening section of level 3/4. The mean of correlation among four categories was .54, indicating a modest level of relationship.

Table 3
Correlation Coefficient Between the Predicted and the Real Mean Score

	Test A		Test B	
	Listening	Reading	Listening	Reading
Pearson Correlation Coefficient	r = .491	r = .579	r = .569	r = .534

The overall results imply that the item writers perceived the test takers' range of language ability as being higher than what was demonstrated by the level description of the test. Since the test was originally developed as a contest with the purpose of discriminating test takers' ability in order to rank the test takers and give prizes, item writers increased the difficulty of each item and intentionally set the test takers' language ability higher than the average. In other words, the nature of a contest affects item writers' view toward the

minimally accepted ability of test takers. To avoid the result of facing several first ranking prizewinners, item writers intentionally made very difficult items. It may have been a contributing factor to the 19 points of score gap between the predicted and the real scores.

As Downing (2006) points out cut scores in general are somewhat arbitrary because they are set up based on human judgment. Because of this trait, it may be said that cut scores cannot be used for the purpose of making institutional decision or scoring interpretation. On the other hand, Popham (2003) argued that when cut scores are established through systemic and scientific research supported by a strong theoretical background, they can be considered reliable parameters. In other words, to measure the validity of a test, the process of investigation to compare the amount of correlation between the real rates and the predicted rates of test scores is needed.

In case of the cluster analysis, it is useful to check how many clusters can be identified by visual presentation. It can be obtained by using Ward Method from SPSS (Statistical Package for Social Science) software. In Figure 1, two dendrograms are presented which were obtained by using Ward Method for Test A and Test B. As figure 1 illustrates there were two clustered test items for the 60 items on Test A. Two clustered test items were also identified for the 60 items on Test B. The number of clusters can be identified by checking the equally long horizontal lines in the rescaled distance cluster combine. The results indicate that the test writers estimated that two levels could be identified by using Test A for the test takers. They also believed that two levels could be set up in case of Test B. It means that visually there is one pass/fail boundary.

For further analysis, Ward method and the K-means method were carried out. It is a good way of securing the reliability of the results to compare the results of different methods of cluster analyses (Scholfield, 1998). Therefore, the two methods were used to corroborate the matches in cluster membership. In case of the items on Test A, the level of agreement between the two methods was 100%. It confirms that one pass/fail boundary was identified by the test writers' predicted scores. Eight items belonged to one cluster and fifty two items belonged to the other cluster.

In case of the items on Test B, the level of agreement between the two methods was measured at 93%. Among 60 items, 4 items were categorized differently between the two methods. In both methods, ten items belong to one cluster and fifty items belong to the other cluster. It means that one pass/fail boundary was also identified by the test writers' predicted scores for Test B.

The results of cluster analyses support that the test writers assumed the one cut score was suitable for each test set. Ideally, it is possible to obtain an expected cut score, i.e. pass and fail boundary. According to the cluster membership, analysed by the test writers' predicted scores, the expected cut score for Test A is 63.75, which is identified by the eighth rank among the 60 mean scores of the items.

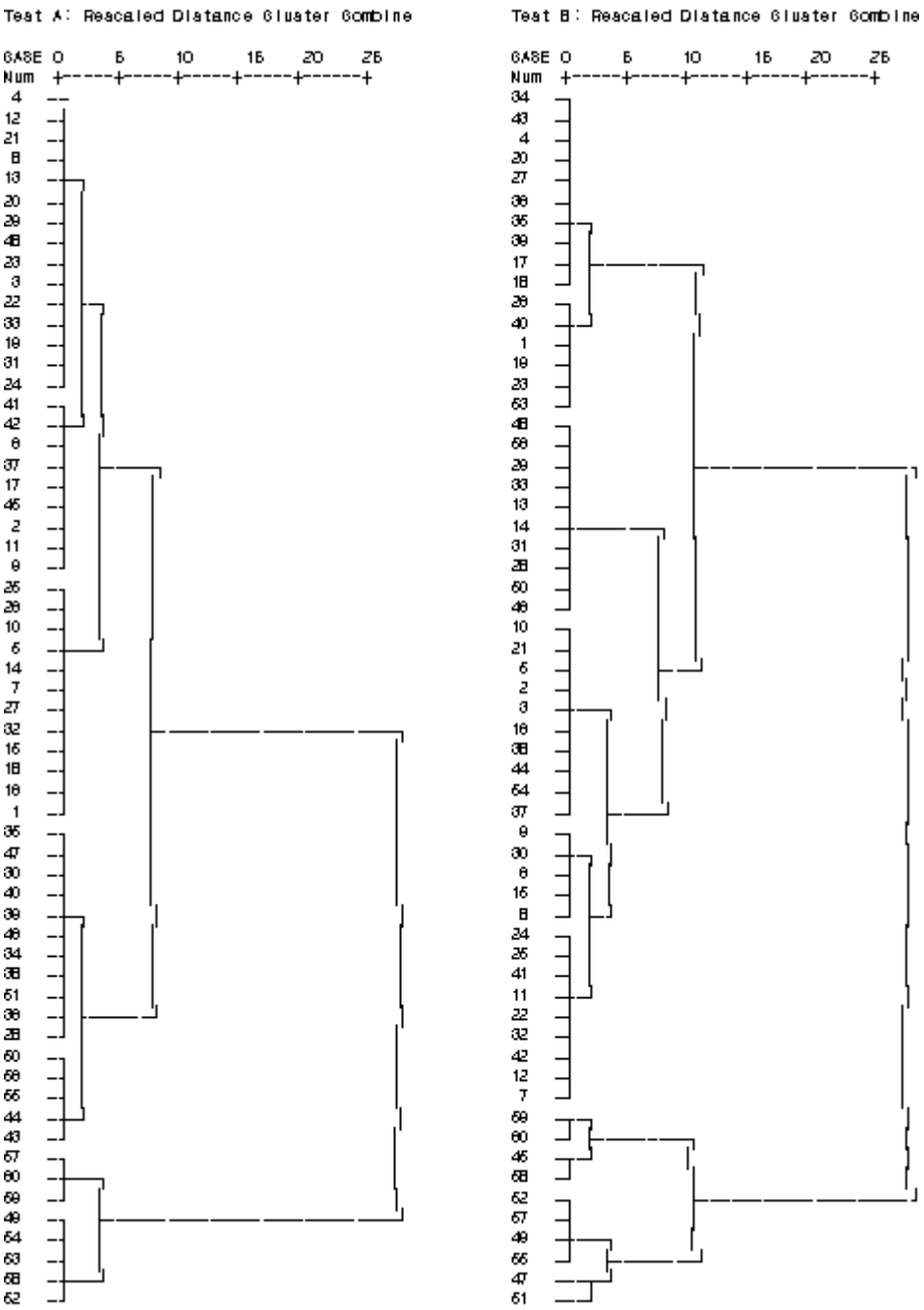

Figure 1. Dendrogram using ward method: Test A and Test B.

In case of the learners' cut score[2], the eighth rank mean score is 26.69, showing a 37.06 points of score gap. In case of Test B, the expected cut score, identified by the tenth rank, is 75.00 and the learners' cut off score is 37.32, showing a 37.68 points of score gap.

All in all, the comparison between predicted cut score and the cut score of test takers shows the fact that item writers intentionally set test takers ability higher than what the real scores show when they wrote test items. This was caused by the characteristic of the test. The test was originally created for a contest that put the test takers in a linear ranking to bestow prizes. However, to enhance the validity of the test as a standardized proficiency test, the gap between the predicted cut sore and real cut score should be narrowed since the focus of the test will be on measuring learners' overall proficiency.

In analyzing the data, we may notice the gap between the expected and the real cut scores. Therefore, it justifies the need of an empirical study checking the validity of pre level classification by utilizing the collected data after the implementation of the test. This kind of effort will provide useful information for test writers. Test developers may use the methods used in the present study.

V. CONCLUSION

Most tests have a role of making decisions. Organizations such as professional associations, academies, and boards of education are all making decisions on the basis of sound information derived from the results of a test. Because of the importance of this role, the process to set up a cut score should be organized in deliberate, considered and defensible manner. Therefore, the empirical and theoretical research of the process is required to enhance the validity of a test. In this sense, validation of a test is an essential part of the development process. Without assessing test validation, one cannot know if a test is adequately qualified in test score interpretation and use. The aim of this study was to illustrate how to set up a pass/fail boundary, i.e. a cut score, for a standardized test for young learners in a valid way.

The data from 1,378 test takers and eight test developers were utilized for the analysis. To get an expected cut score, the Angoff method was used and it was compared with the actual score obtained from the test. The mean of correlation coefficients between two scores was .54, indicating a moderate level of correlation. Cluster analyses based on the

[2] Theoretically, it is possible to conduct a cluster analysis to identify a pass/fail boundary in a valid way by using the learners' raw data. However, due to an ethical reason (the confidentiality of the individual learners' exam data), it cannot be presented in the present study.

test writers' data were conducted to identify the number of possible levels. The results indicate that there were two clustered memberships, indicating one pass/fail boundary. However, there was a noticeable gap between the expected cut score and the actual cut score.

The results of the present study suggest that if the items on a test are reliable, then using the test writers' predicted score for English proficiency tests should help to identify the pass/fail boundary or cut score. Therefore, it is necessary to check the correlations between test writers' predicted scores and the real scores. The results may help the test writers to anticipate the learners' level in a more reliable way, and it may eventually enhance the reliability and validity of a consequent English proficiency test. The results of cluster analyses in the present study indicate that it is a possible procedure to determine the pass/fail boundary through statistical analysis. In other words, to obtain a high level of validation, empirical data should be obtained and statistically tested.

It would be ideal if we could present the results of factor analysis on the test items to confirm the construct validity of the test. Also, it would be helpful for the readers if we could provide the results of item analysis, and item difficulties and discriminations of individual test items. However, because of the confidentiality of the results of those analyses—which is often the case in studies on testing—and the limitation of the space in this article, those were not covered. It can be cautiously mentioned that the overall results of those analyses were acceptable for further analysis of cluster analysis in the present study. Also, it would be ideal if we could identify cut off scores of speaking and writing. Unfortunately, those skills were not systematically assessed in this study. Therefore, checking speaking and writing cut-off scores were not covered in the present study. We suggest that identification of cut off scores of productive skills should be carried out in further studies.

To improve overall Korean English language education, qualified English language tests have to be developed. To do that, validation research on tests has to be implemented since information gained from research is important in the development of valid tests. In Korea some testing tools developed by foreign testing service organization have been widely used as major English language tests. However, since those standardized tests originally developed for test takers who are trying to study at universities in North America or being engaged in international trade, it is questionable whether they can accurately measure Korean test takers' English language proficiency. In this sense, the construct validity of these tests may be low for Korean test takers. Considering the importance of developing testing tools for Korean test takers (Hyun-Ju Kim, 2009), efforts are being made to design a National English Ability/Proficiency Test (NEAT/NEPT). In developing level classifications in NEAT/NEPT, the methods used in the presented study may be referred to. In addition to research on validation, to enhance the validity of standardized test items,

another area to be studied is the effectiveness of systematized training for test writers. The training of test writers is a long-term issue that would assure the validation of not just one test but a whole series of tests. Therefore, issues in the training of test item writers have to be investigated in further research.

REFERENCES

Bachman, L. F., & Palmer, A. S. (2000). *Language testing in practice*. Oxford: Oxford University Press.

Chapelle, C. A., Jamieson, J., & Hegelheimer, V. (2003). Validation of a web-based ESL test. *Language Testing, 20*(4), 409-439.

Cizek, G, J. (2006). Standard setting. In S. M. Downing, & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 225-258). Mahwah, NJ: Lawrence Erlbaum.

Downing, S. M. (2006). Twelve steps for effective test development. In S. M. Downing, & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 3-25). Mahwah, NJ: Lawrence Erlbaum.

Gottlieb, M. (2003). *Large-scale assessment of English language learners: Addressing educational accountability in K-12 settings.* Alexandria, Virginia: Teachers of English to Speakers of Other Languages.

Hasselgreen, A. (2000). The assessment of the English ability of young learners in Norwegian schools: An innovative approach. *Language Testing, 17*(2), 261-277.

Hughes, A. (2003). *Testing for language teachers*. Cambridge: Cambridge University Press

Jung, Hee-Jung. (2008). *A study of predicting the variables of difficulty levels on middle school students' English reading proficiency test items.* Unpublished doctoral dissertation, ChungAng Unviersity, Seoul.

Keyser, D. J., & Sweetland, R. C. (1984-1994). *Test critiques* (No. 1-10). Kansas City: Test Corporation of America and Austin.

Koo, So-Young, & Kim, Hae-Dong. (2007). Analysis of a test for primary school English learners and level descriptions. *Primary English Education, 13*(3), 147-176.

Kim, Hyun-Ju. (2009). An antecedent study of the state-wide English language ability test. *Modern English Education*, *10*(2), 44-59.

Lee, Young-Shik, & Kim, Hye-Young. (2009). A comparative study of the achievement standards between the Revised Korean National Curriculum of English and Common European Framework of References (CEFR). *Modern English Education*, *10*(2), 108-132.

McKay, P. (2006). *Assessing young language learners*. Cambridge: Cambridge University

Press.

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

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103). New York: Macmillan.

Popham, W. J. (2003). Seeking redemption for our psychometric sins. *Educational Measurement: Issues & Practice, 22*(1), 45-48.

Romesburg, H. C. (2004). *Cluster analysis for researchers*. Florida: Krieger.

Scholfield, P. (1998) *Multivariate statistics: Cluster analysis*. Unpublished manuscript, Essex University.

Seong, Tae-Jae. (2005). *Principles and applications of Item response theory*. Seoul: Kyo-yuk Kwa-hak-sa.

Shin, Dong-Il. (2002) The development of descriptors on scales of L2 productive skills utilized by the many-faceted Rasch model. *English Teaching, 57*(4), 469-499.

Shin, Dong-Il, Song, Ji-Hyun, & Kim, Na-Hee. (2006). Examining English language assessment for young learners in Korea: Focus on level descriptors and item types. *Foreign Languages Education, 13*(4), 243-269.

Weir, C. J. (2005). *Language testing and validation*. Basingstoke, UK: Palgrave Macmillan.

Haewon Pyo
TESOL department
Hankuk University of Foreign Studies
270 Dongdaemun-Ku, Immun-dong
Seoul, Korea 130-791
Tel: 02-2173-3017
Email: hw-pyo@hanmail.net

Haedong Kim
ELT, Graduate School of Education
Hankuk University of Foreign Studies
270 Dongdaemun-Ku, Immun-dong
Seoul, Korea 130-791
Tel: 02-2173-3017
Email: khd@hufs.ac.kr

Received 22 December 2010
Revised 4 February 2011
Accepted 8 February 2011