THE JOURNAL OF ASIA TEFL
Vol. 15, No. 4, Winter 2018, 1021-1035
http://dx.doi.org/10.18823/asiatefl.2018.15.4.9.1021

The Journal of Asia TEFL

http://journal.asiatefl.org/

e-ISSN 2466-1511

Evaluating the Validity of Classroom Assessments in the Secondary EFL Curriculum

Hyun Jung Kim
Hankuk University of Foreign Studies, Korea

Sung Su Yang
Sung-ho High School, Korea

This study investigated the validity of instructionally relevant classroom assessments embedded within a secondary English as a Foreign Language (EFL) curriculum. The study was conducted at a high school in Korea, including 148 12th graders learning EFL. For the validity analysis, two sets of achievement tests were chosen (i.e., multiple-choice listening tests for the midterm and final exams). The test items were qualitatively analyzed in relation to the national curriculum and instructional materials, and student test performance was also statistically analyzed. Results revealed that the two achievement tests were found to tap all curricular learning goals, albeit in an imbalanced manner due to the varied degree of coverage of the learning goals (cognitive validity). While the tests covered the target instructional contents, the degree of coverage and representativeness was again limited and inconsistent with the instructions (instructional validity). A statistical analysis of the test performance provided evidence of test validity but found a lack of consistency between the students' listening ability and item difficulty levels (inferential validity). These findings suggest the types and forms of evidence required for test development and a validity argument within a learning-focused classroom assessment context.

Keywords: classroom assessment, validation, validity, listening test, achievement test

Introduction

Learning-focused classroom assessments have recently begun to gain attention in the second language (L2) testing field, informed by research in general education. L2 assessment researchers have recognized the central role of assessment for language learning in the classroom and have refined the definition of Learning-Oriented Assessment (LOA) for L2 educational contexts (Colby-Kelly & Turner, 2007; Turner, 2012; Turner & Purpura, 2016). LOA represents an attempt to promote L2 learning by bridging L2 instruction and assessment (e.g., Green & Hamp-Lyons, 2015; Turner & Purpura, 2016). LOA emphasizes the learning aspects of assessment, putting more emphasis on the learning versus measurement elements; therefore, it promotes assessment *for* learning.

Considering the nature of LOA, in which instruction, learning, and assessment are interconnected, there is a growing body of theoretical and empirical research on L2 classroom assessment. These prior studies have reported the development and effectiveness of classroom assessment (e.g., Colby-Kelly, 2014; Kim & Kim, 2017; Tsagari, 2014). Tsagari (2014) explored teachers' use of unplanned assessment in the EFL classroom and discussed the complexities involved in implementing learning-focused assessment in

classrooms. Likewise, Kim and Kim (2017) explored the nature of reading-to-write tasks used in an English for Academic Purpose (EAP) context and examined how such tasks could be used for learning in a natural classroom setting.

Despite growing interest in learning-focused assessments, research on L2 classroom assessment is still under way, and findings have been far from conclusive. While prior studies tend to focus on the development/implementation and effectiveness of L2 classroom assessment, there has been a lack of research regarding its validity. In order to evaluate the usefulness of L2 classroom assessment, there is a need for research on how adequately and appropriately L2 classroom assessment scores are interpreted and used. Despite its importance, little attention has been devoted to the validation of instructionally relevant L2 classroom assessments (Pellegrino, DiBello, & Goldman, 2016).

The proposed research aims to investigate the validity of classroom assessments in the Korean high school context, where L2 learning and teaching are controlled by the National Curriculum provided by the Ministry of Education. The research uses Pellegrino et al.'s (2016) validity analysis framework, which has been proposed as conceptualizing "the multiple components of validity applicable to assessments intended to function at the classroom level to support ongoing processes of teaching and learning" (p. 60). It consists of three components: cognitive validity, instructional validity, and inferential validity. Since these three components (individually and jointly) are critical for examining the validity of classroom-based assessment that is intended to enhance learning and instruction (Pellegrino et al., 2016), the current research evaluates the validity of learning-focused EFL classroom assessments from an individual and collective analysis of the three validity components.

Literature Review

Before examining the validity of instructionally relevant classroom assessments, the unique characteristics of classroom assessment and prior research on it are first reviewed. A review of the theoretical framework used for the validation of classroom assessment also provides a background for the validity analysis of the current research.

Classroom Assessment

Classroom-based assessment refers to an assessment that is internal to the classroom, and is oftentimes managed by the teacher (Turner, 2012). It has long been believed that classroom-based assessment is no different from traditional large-scale tests in terms of its development and use, which are external to the classroom (e.g., standardized tests). That is, test tasks/item types used in large-scale testing have been used in classroom-based assessment, and test results have been interpreted and used in the same way as large-scale testing (Turner, 2012). Recently, however, L2 assessment researchers have begun to recognize the unique characteristics of classroom-based assessment, which has led to a growing interest in assessment focusing on teaching and learning in the L2 testing literature (Leung, 2004; Rea-Dickins, 2006). Classroom-based assessment shares the same fundamental features as general language assessment by including "a systematic procedure for eliciting test and nontest data (e.g., a teacher checklist of student performance) for the purpose of making inferences or claims about certain language-related characteristics of an individual" (Purpura, 2016, p. 191). However, such assessment further involves teachers' planning and collection of multiple forms of information regarding students' language ability and use, most often in relation to regular instructional activities. Moreover, it focuses not only on the interpretation of students' language ability/progress, but also on additional teaching and learning as a result of the assessment (Colby-Kelly & Turner, 2007; Yin, 2010). Considering its unique nature, the principles of development and the use of classroom-based assessment are first introduced before discussing its validation.

Classroom-based assessment is often administered as a form of achievement test used to measure the

degree to which students have learned or mastered content within a specific instructional domain (Bachman & Palmer, 1996; Brown, 2005; Hughes, 2003). Examples include a final achievement test, which is given at the end of a course or program, and a progress achievement test, which measures students' progress during the course of study (Hughes, 2003). Since the purpose of testing is closely related to a specific instructional domain, test content is context-dependent, meaning that it is based on the objectives of the course or program. That is, test content includes the skills/knowledge learned and practiced in the classroom or the course objectives presented in the instructional materials or syllabus. The test is developed based on classroom teaching and learning, and the test results are in turn used to improve subsequent teaching and learning by making changes in the course design and materials, in addition to making decisions for advancement or graduation.

Another key feature of an achievement test is the nature of the criterion-referenced test. According to Brown (2005), criterion-referenced tests measure well-defined, objective-based language points to evaluate the amount of material that students have learned. Due to this nature, students oftentimes already know what test content to expect; thus, the distribution of test scores is non-normal, and all students can receive a high score if they have studied/mastered the material, regardless of comparisons to other students' performance.

Research on classroom-based assessment has been conducted with respect to various topics. To list a few, researchers have examined (1) different types of information that teachers collect/use to assess students' L2 ability (e.g., Butler & Lee, 2010; Leung & Scott, 2009; Rea-Dickins & Gardner, 2000), (2) alternative types of assessment methods (e.g., Cheng, Rogers, & Hu, 2004), and (3) teachers' assessment and decision-making process and their variability (e.g., Brindley, 2001; Leung & Lewkowicz, 2006; Yin, 2010). Even though research interest and the importance of classroom-based assessment have recently grown in the L2 testing field, there are still various areas to be investigated. Further studies on classroom-based assessment for a wide range of levels (e.g., primary school level to tertiary level) might lead to a better understanding of how teachers develop and use classroom assessment. Research on the quality criteria for classroom-based assessment, such as the reliability and validity of assessment, is also limited in the L2 testing literature (Pellegrino et al., 2016; Turner, 2012). The validation of classroom-based assessment is further expected in developing a solid argument for the development and use of classroom assessment in an instructional context.

Validation of Classroom Assessment

Validity is one of the test qualities that provide evidence of test usefulness (Bachman & Palmer, 1996). Validity is defined as "the meaningfulness and appropriateness of the *interpretations* that we make on the basis of test scores" (Bachman & Palmer, 1996, p. 21), while validation refers to the process of providing evidence of test validity. The notion of validity has evolved over time. Most recently, Kane's (2006, 2012) argument-based approach to validation is widely used in the L2 testing literature. It consists of two components of interpretive and validity arguments. An interpretive argument refers to "the proposed interpretations and uses of assessment results by laying out a network of inferences and assumptions leading from the observed performances to the conclusions and decisions based on the assessment scores," whereas a validity argument "provides an evaluation of the interpretive argument's coherence and the plausibility of its inferences and assumptions" (Kane, 2012, p. 8). Following this argument-based approach, many previous studies on L2 assessment validity have been conducted, mostly for large-scale, and oftentimes high-stakes, tests, whereas little attention has been devoted to the validation of instructionally relevant classroom assessments (Pellegrino et al., 2016).

For the validation of classroom-based assessment, Pellegrino et al. (2016) proposed a validity analysis framework, considering the nature of assessment for instructional settings. This framework encompasses three components, including cognitive, instructional, and inferential validity. First, cognitive validity addresses the extent to which the knowledge, skills, and abilities (KSAs) that learners are supposed to use correspond with those that they actually use in the assessment. Instructional validity addresses the extent

to which an assessment is aligned with instruction and KSAs, as defined by the curriculum. Lastly, inferential validity is concerned with the extent to which an assessment provides reliable and accurate information about learner performance. According to Pellegrino et al. (2016), these three components (individually and jointly) are critical in examining the validity of classroom assessment. In addition, they stressed the importance of contextual factors for the classroom assessment validity argument by stating that "one of the major challenges associated with examining the validity of assessments intended to function close to instruction is that the judgments are tightly coupled to contextual factors related to a student's curricular and instructional experiences" (Pellegrino et al., 2016, pp. 67-68). Therefore, it seems important to understand the curriculum, instruction, and assessment together when examining the three types of validity with respect to L2 classroom assessment. So far, however, there have not been attempts to examine the validity of L2 classroom assessment based on the framework conceptualized for instructionally relevant assessments, considering contextual factors. An application of such a framework might not only enable the systematic validation of L2 classroom assessment, but also might deepen and broaden insights into the development and use of L2 classroom assessment.

The Current Study

The current study examines the validity of classroom assessments embedded within a high school EFL curriculum. It focuses on the validity of the listening section of high school English classroom assessments from an individual and collective analysis of the cognitive, instructional, and inferential validity components. This study addresses the following research questions:

1. To what extent do the KSAs required in the listening assessment correspond to the KSAs, as defined by the curriculum?
2. To what extent are the KSAs required in the listening assessment aligned with classroom instructions?
3. To what extent does the listening assessment provide reliable and accurate information regarding high school students' listening performance?

Methods

For the classroom assessment validity argument, the participants and contextual factors are presented in this section, including the curriculum, instruction, and assessment.

Participants

The study was conducted at a high school located in Korea. The participants included 148 12th grade students in a public high school. Their overall listening ability level was moderately high. This finding comes from the results of the nationwide listening comprehension test, which is jointly administered twice a year by the 15 municipal/provincial Offices of Education across the country for the purpose of improving secondary school students' listening ability. The test is equivalent to the College Scholastic Ability Test (CSAT), which is a standardized test used for college entrance in Korea, in terms of the test structure and the types/difficulty of test items. The participants' average test score on the test (targeting 12th graders) was 14.8 out of 20. In addition to English lessons at school, most participants were taking online and offline private English lessons to prepare for the school tests and the CSAT.

The four intact classes of 148 students were taught by one non-native, female English teacher in her middle 40s with a master's degree in TESOL. She had been teaching at secondary public schools for 17 years at the time of the data collection. She was in charge of teaching the listening and speaking parts of the English lessons, meeting with each intact class three times a week, while another English teacher was

responsible for the reading and writing parts of the English lessons for the same group of students.

Curriculum

The curriculum used in the current study was the 2009 Revised National Curriculum (Ministry of Education, Science, and Technology, 2009). The National Curriculum provides the goals/objectives for each skill, teaching/learning methods, and evaluation goals and tips for each type of English textbook. Considering the textbook used in the classroom, the main focus of the curriculum was the listening section of the English Conversation textbook in the current study.

High school listening in the curriculum largely aims at mastery of four skills, including (1) understanding the main idea (topic/gist of a speech); (2) finding the details (detailed information in a speech); (3) finding the logical relationships (relationships among the ideas presented in a speech); and (4) understanding inferences (implicitly stated information such as the speaker's intention, attitude, and implications). More concrete curricular learning goals introduced in the curriculum (Ministry of Education, Science, and Technology, 2009, pp. 292-293) and the accompanying research (Korea Institute for Curriculum and Evaluation, 2012, pp. 330-333) can be described as follows.

TABLE 1
Curricular Learning Goals

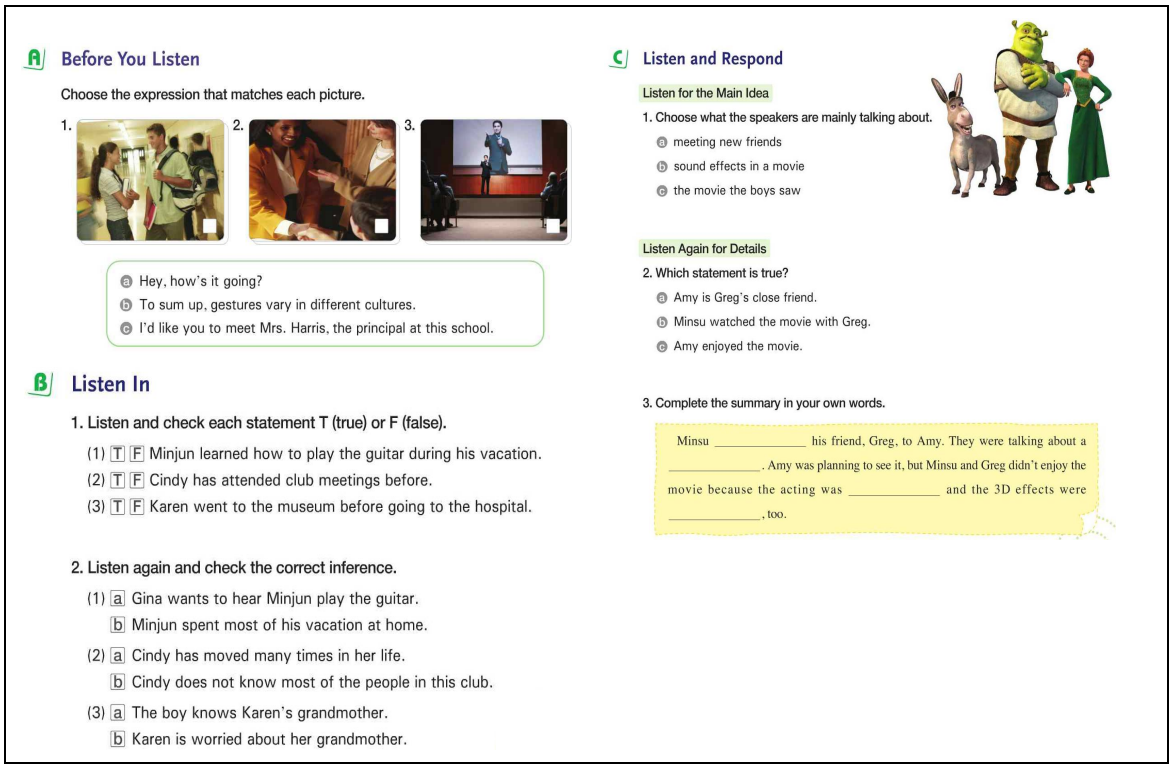
Target skill	Curricular learning goals	
Main idea	Listeners can understand the topic or gist of a speech or oral discourse on a range of topics.	• Listeners can understand the topic of a speech or oral discourse on a range of topics. • Listeners can understand the gist of a speech or oral discourse on a range of topics.
Detail	Listeners can understand the important details in a speech or oral discourse on a range of topics.	• Listeners can understand important details in the types of speech or oral discourse that involve explanations on a range of topics. • Listeners can understand important details in the types of speech or oral discourse that involve discussions on a range of topics.
Logical relationship	Listeners can understand the logical relationships among ideas within a speech or oral discourse on a range of topics.	• Listeners can understand the sequence of events in a speech or oral discourse that involve process or procedures. • Listeners can understand the causes and effects in a speech or oral discourse on a range of topics.
Inference	Listeners can understand implicit and inferred information, including the speaker's intention and main purpose of the discourse. Listeners can understand the speaker's tone and attitude in a speech or oral discourse on a range of topics. Listeners can understand the implications made in a speech or oral discourse on a range of topics.	• Listeners can understand the purpose of a speech or oral discourse on a range of topics. • Listeners can understand the speaker's intention in a speech or oral discourse on a range of topics. • Listeners can understand the speaker's tone in a speech or oral discourse on a range of topics. • Listeners can understand the speaker's attitude in a speech or oral discourse on a range of topics. • Listeners can understand the differences in opinion in a speech or oral discourse that present differing views on a range of topics. • Listeners can understand the moral or implications of a fable or anecdote on a range of topics.

Materials

Two types of materials (i.e., textbook and supplementary material) were used in class for listening practice. Based on the National Curriculum, publishers develop different types (e.g., English Conversation, Reading and Writing, Practical English, Advanced English Reading) and levels of textbooks (i.e., for each grade of middle and high school) in Korea. The Ministry of Education approves

certain textbooks on the basis of an evaluation. Each school then selects and uses one among many textbooks approved by the Ministry of Education.

The high school of the current study selected the textbook, *High School English Conversation* (Ahn, Oh, Kim, Choi, & Kim, 2013), for the listening and speaking parts of the English lessons. The textbook consists of eight units, including eight different topics (e.g., good manners, women and men, teen worries, and hopes for the future); additionally, each unit includes various listening and speaking activities. With respect to the listening part, for example, the textbook introduces diverse activities such as matching, true/false, and listening for the main idea/details, and a summary, as seen in Figure 1 below.



A Before You Listen

Choose the expression that matches each picture.

1. 2. 3.

ⓐ Hey, how's it going?
ⓑ To sum up, gestures vary in different cultures.
ⓒ I'd like you to meet Mrs. Harris, the principal at this school.

B Listen In

1. Listen and check each statement T (true) or F (false).
(1) T F Minjun learned how to play the guitar during his vacation.
(2) T F Cindy has attended club meetings before.
(3) T F Karen went to the museum before going to the hospital.

2. Listen again and check the correct inference.
(1) a Gina wants to hear Minjun play the guitar.
b Minjun spent most of his vacation at home.
(2) a Cindy has moved many times in her life.
b Cindy does not know most of the people in this club.
(3) a The boy knows Karen's grandmother.
b Karen is worried about her grandmother.

C Listen and Respond

Listen for the Main Idea

1. Choose what the speakers are mainly talking about.
ⓐ meeting new friends
ⓑ sound effects in a movie
ⓒ the movie the boys saw

Listen Again for Details

2. Which statement is true?
ⓐ Amy is Greg's close friend.
ⓑ Minsu watched the movie with Greg.
ⓒ Amy enjoyed the movie.

3. Complete the summary in your own words.

Minsu _____ his friend, Greg, to Amy. They were talking about a _____. Amy was planning to see it, but Minsu and Greg didn't enjoy the movie because the acting was _____ and the 3D effects were _____, too.

Figure 1. Example listening activities (Ahn et al., 2013, pp. 14-15).

In addition to the textbook, other supplementary listening material (Educational Broadcasting System, 2017) was used in class for listening practice and preparation for the CSAT due to the limited number of listening activities in the textbook. The supplementary material includes speeches and discourse on a range of topics such as school life, culture, hobbies, and travel. The listening texts are also accompanied by multiple-choice questions, which address various types of listening skills required in the curriculum (understanding the main idea, finding the details, finding the logical relationships, and understanding inferences). Using the textbook and supplementary material, students practiced English listening three times a week, 50 minutes per each lesson.

Instruments

The teacher developed and administered two periodic assessments in order to evaluate the extent of learning after the listening practice in class: a midterm and a final. For each assessment, 17 multiple-choice items were given with five options (one key and four distractors), which measured four target skills (understanding the main idea, details, logical relationships, and inferences), according to the teacher's test specifications. (The constructs measured in each individual item are presented in the Results and Discussion section below.) Due to the nature of achievement tests, the teacher used the same listening passages used in class for the development of the tests; however, all of the test items were changed. That

is, different skills were required to answer the items. For the purpose of the analysis in this study, when the answer was correct, one point was given, for a maximum total of 17 points.

Data Collection Procedures

To evaluate the validity of the classroom assessments, three types of data were collected. In order to understand the KSAs that the high school students are supposed to learn in terms of English listening, the Revised 2009 National Curriculum was used. As mentioned above, the listening section of the English Conversation textbook of the curriculum became the main focus. The second type of data collected was the English Conversation textbook and the supplementary listening material used for classroom instruction. Only the chapters covered before the midterm and final were collected, respectively. The last type of data was the actual midterm and final tests, with the listening scripts and students' responses to each test item. All three types of data were provided by the teacher without any identification number or student name.

Data Analysis

The three types of data were analyzed individually and compared to one another in order to examine the three components of validity for instructional settings (cognitive, instructional, and inferential validity). Basically, the analytical procedure was followed by Pellegrino et al.'s (2016) framework. To examine the cognitive validity (Research Question 1), the types and levels of knowledge tapped by the midterm and final items and the curricular learning goals were compared. For this comparison, each test item on the midterm and final tests was coded on the basis of the target skills and learning goals of the curriculum by the researcher and another English teacher with extensive experience in teaching English at public secondary schools. For the analysis of instructional validity (Research Question 2), the two coders first classified each activity and question in the textbook and the supplementary material into subskills of the target skills specified in the curriculum (Table 1) so as to identify what had been taught during the instruction. On the basis of this classification of subskills, the coders then coded each individual test item to examine the correspondence between the assessment and the classroom instructions (i.e., targeted skills in the instruction vs. assessed skills on the tests). Lastly, to examine the inferential validity (Research Question 3), student performance was statistically analyzed for the midterm and final tests. These statistical analyses included descriptive statistics, a test reliability analysis, and an analysis of a dichotomous Rasch model using the FACETS program. They examined the appropriateness of the midterm and final tests, and further analyzed the reliability and accuracy of the interpretation of students' listening test performance.

Results and Discussion

The findings of the analysis are presented for each component of validity for instructional settings (cognitive, instructional, and inferential validity) by addressing each research question.

Analysis of Cognitive Validity

For the analysis of cognitive validity, the extent to which the test items correspond to the curricular learning goals was analyzed (Research Question 1). Table 2 below summarizes the comparison of the target skills/curricular learning goals and the test items measuring a relevant goal. The percentages of each target skill measured on the midterm and final are presented with the actual test item numbers.

TABLE 2
Analysis of Cognitive Validity

Target skill	Curricular learning goals	Midterm	Final
Main idea	Listeners can understand the topic or gist of a speech or oral discourse on a range of topics.	8, 11, 14 **(17.6%)**	1, 8, 11, 15 **(23.5%)**
Detail	Listeners can understand the important details in a speech or oral discourse on a range of topics.	2, 3, 5, 6, 7, 9, 10, 12, 13, 15 **(58.8%)**	2, 3, 5, 6, 7, 9, 10, 12, 13, 14 **(58.8%)**
Logical relationship	Listeners can understand the logical relationships among ideas within a speech or oral discourse on a range of topics.	16, 17 **(11.8%)**	16, 17 **(11.8%)**
Inference	Listeners can understand implicit and inferred information, including the speaker's intention and main purpose of the discourse. Listeners can understand the speaker's tone and attitude in a speech or oral discourse on a range of topics. Listeners can understand the implications made in a speech or oral discourse on a range of topics.	1, 4 **(11.8%)**	4 **(5.9%)**

Overall, both the midterm and final tests covered all four target skills and the curricular learning goals of the curriculum. However, the degree of coverage varied across the target skills, and the pattern of the coverage was similar between the two tests. For example, most of the test items on both tests focused on understanding details (58.8% for each test), while only a couple of logical relationship (11.8% for each test) and inference items (11.8% and 5.9%, respectively) appeared on the tests. In particular, inference skills require three different aspects of listening comprehension in the curriculum, including understanding (1) implicit and inferred information; (2) the speaker's tone and attitude; and (3) implications. However, both tests included only the first aspect of inference (implicit and inferred information) by asking about the intended purpose of the speaker's announcement (midterm item 1), the place where the conversation might be taking place (midterm item 4), and the relationship between the speakers inferred from the conversation (final item 4). As a result, the tests assessed students' ability to infer the meaning of a speech or oral discourse within a limited context, thereby providing somewhat limited evidence of cognitive validity.

Considering the curricular learning goals, which equally emphasize the four target skills (Table 1), a lack of inference items regarding diverse aspects seems to lower the cognitive validity of the midterm and final tests. Inference skills are a newly introduced component from high school English in the curriculum, which are not included in elementary and middle school English; more specifically, different aspects of inference are added only for the second and third grade levels of the high school curriculum, while a limited component of inference (implicit and inferred information) is included in the curriculum for the first grade level (Ministry of Education, Science, and Technology, 2009). In other words, listening practice to infer meaning in a wide range of contexts is a unique characteristic of upper-level high school English; however, such an opportunity was not provided for both assessments. To sum up, the midterm and final tests did not represent the curricular learning goals in a balanced manner, even though both tests tapped all four target skills, as identified in the curriculum.

Analysis of Instructional Validity

For the analysis of instructional validity, the degree to which the assessed skills on the tests correspond to those practiced in the classroom instructions was examined (Research Question 2). Before this comparison, the skills that the students practiced during the instructions were initially identified, which yielded a total of 15 targeted skills in the instructions. They were all related to the four curricular learning goals to some extent; thus, it appears that the instruction was prepared and delivered in relation to the curriculum. For example, students practiced understanding the different types of main ideas (*Listen to a monologue and identify its purpose; Listen to a dialogue and identify the speakers' opinions*). They also

practiced finding details (*Listen to a dialogue and understand information about numbers; Listen to a dialogue and find matching content/information in a chart*). For understanding logical relationships, students listened to a dialogue and found what the speaker has to do/what the speaker asks someone to do; they also identified logical reasons. Lastly, students identified an appropriate response by making an inference after listening to a short/long dialogue. More concrete targeted skills practiced in the instructions are listed in Table 3, following the instructional sequence. These 15 skills were practiced prior to both the midterm and final tests; however, the instructions up to the midterm test focused more on skills from 1 to 8, while the focus up to the final test was on skills from 9 to 15.

Table 3 presents the results from the analysis of each test item in comparison with the targeted skills in the instructions. The left column represents the targeted skills, whereas the right two columns indicate the accompanying items assessed on the midterm and final tests.

TABLE 3
Comparisons between Targeted Skills in the Instruction and Assessed Skills on the Tests

Targeted skill	Assessed skill	
	Midterm	Final
1. Listen to a monologue and identify its purpose.	1	1
2. Listen to a dialogue and identify the speakers' opinions.	2	
3. Listen to a monologue or dialogue and identify the topic.	8, 11, 14	8, 11, 15
4. Listen to a dialogue and identify the relationship and tone of the speakers, and where the conversation is taking place.	4	4
5. Listen to a dialogue and identify information that is inconsistent with the pictures.		
6. Listen to a dialogue and understand what the speaker has to do, and what the speaker asked someone to do.	5	5
7. Listen to a dialogue and identify the reasons.	6	6
8. Listen to a dialogue and understand information about the numbers (e.g., the amount of payment).		
9. Listen to a dialogue and identify information/content that is not mentioned.	12, 13	13
10. Listen to a monologue and identify information that is inconsistent.	3, 7, 9, 10	3, 7, 10, 12
11. Listen to a dialogue and find matching content/information in a chart.	15	2, 9, 14
12. Listen to a short dialogue and select an appropriate response.		
13. Listen to a long dialogue and select an appropriate response.	16, 17	16, 17
14. Listen to a monologue and select an appropriate response for the speaker in a given situation.		
15. Listen to a monologue and understand the topic and supporting details.		

The comparisons between the instructions and the tests showed several interesting findings. First, as found in the analysis of cognitive validity, the most frequently assessed skill was detail related. For example, skills 9, 10, and 11 in Table 3, asking about the details of a dialogue/monologue, were often tested on both tests (seven and eight items out of 17 items for each test, respectively). Another interesting feature is a mismatch between the instructions and test items. As previously identified, the focus of the instructions up to the midterm test was on the targeted skills from 1 to 8 in the table, although students briefly practiced other types of skills in class. On the other hand, the focus up to the final test was on the targeted skills from 9 to 15. However, as seen in Table 3, very similar skills were measured on both the midterm and the final, regardless of the focus of the instructions. Another mismatch between the targeted and assessed skills was found in the missing skills on the tests. Students practiced all 15 skills for one semester, but some of the skills never appeared on the tests (e.g., skills 5, 8, 12, 14, 15). While the same skill was repeatedly tested even within each test (e.g., skills 3, 10), some skills were tested only once if they appeared on the tests (e.g., 1, 4, 6, 7). From these comparisons, it seems that the tests were not developed systematically with regard to the instructions. Rather, they appear to align with the CSAT, particularly when considering the mismatch between the instructions and test items. This might have happened because the participants were 12th graders who had already learned most of the target KSAs required in the high school English curriculum, and the CSAT is a high-stakes test that determines college entrance (Chon, 2014; Kwon, Lee, & Shin, 2017). However, the overall mismatch between the

instructions and tests is problematic, considering the nature of achievement tests (e.g., midterm and final tests). These achievement tests aim to measure the extent of learning within a specific instructional domain, in which test content is based on the objectives of the course, textbook, and/or syllabus (Bachman & Palmer, 2010; Brown, 2005; Hughes, 2003). Due to these characteristics, students should be able to know what content to expect on the tests. Since these tests failed to sufficiently reflect the instructional objectives in a balanced manner, there seems to be a lack of validity evidence, which might make it difficult to expect positive washback from the students and might prevent them from acquiring further learning as a result of the assessment (Bachman, 2005; Bachman & Palmer, 1996; Colby-Kelly & Turner, 2007; Pellegrino et al., 2016; Purpura, 2016).

Analysis of Inferential Validity

Lastly, for the analysis of inferential validity, students' test performance was statistically analyzed. For an overview of their test performance, descriptive statistics were first calculated for each of the midterm and final tests. As seen in Table 4, students' performance on both tests was generally good (mean of 12.86 and 11. 50 out of 17) but showed a wide range in listening comprehension (range of 16 and 15). The skewness and kurtosis values were within ±2, indicating that the distributions of the test scores were normal, instead of negatively skewed, although the tests were achievement based (Bachman, 2004). Also, both tests were internally consistent at a moderately high level (Cronbach's α of 0.83 and 0.79), meaning that the items on each test consistently measured the same trait to a great extent (i.e., listening ability).

TABLE 4
Descriptive Statistics of the Midterm and Final

	Midterm	Final
Mean	12.86	11.50
Median	13.00	12.00
SD	3.70	3.61
Skewness	-0.74	-0.50
Kurtosis	-0.25	-0.31
Minimum	1.00	2.00
Maximum	17.00	17.00
Cronbach's α	0.83	0.79

A further Rasch analysis was conducted for each test in order to examine student performance. A 2-facet dichotomous Rasch model (including the examinee and item facets) was analyzed, and Figure 2 below shows the overall results from the analysis of the midterm test performance. The first column represents an equal interval scale, called a logit scale, which enables not only the elements of each facet to be compared with one another, but also enables different facets to be compared on the same scale. The next two columns present each of the examinee and item facets. That is, the students are presented in descending order of the ability measure (the highest performance at the top), and the items are depicted in descending order of item difficulty (the most difficult item at the top).

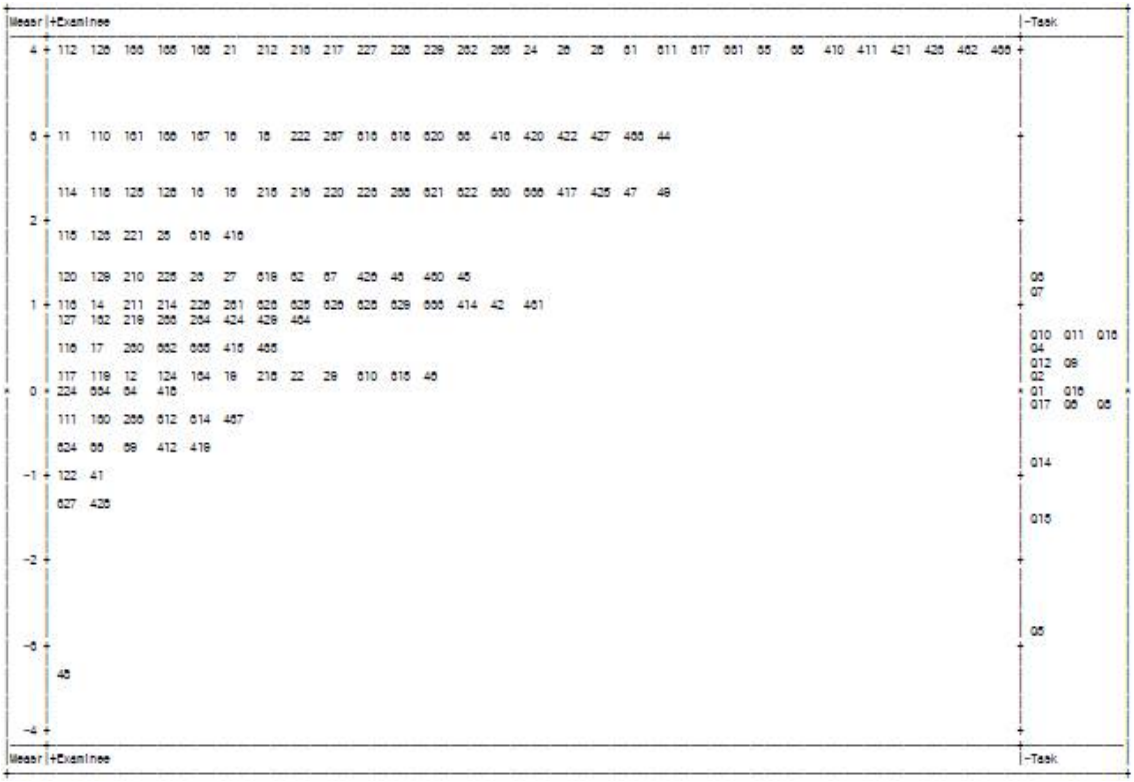


Figure 2. All-facet map for the midterm test.

As seen in the second column of the map (examinee facet), many examinees are located at the middle to the top, indicating that many students received a good score on the test. The measure of examinee ability ranged from -3.27 to 4.32, with a 7.59 logit spread. With the strata value of 2.04 (R = 0.62), this group of students could be separated into two statistically distinct levels of listening ability. Since the separation reliability coefficient (R = 0.62) was not large enough, it seems that there is error variance involved in the examinee ability measures (e.g., other than listening ability) (Myford & Wolfe, 2000). This error variance can be mainly attributed to the weak alignment between students' listening ability levels and item difficulty levels revealed in Figure 2. The students showed a wide range of listening ability, but the midterm test items used to measure the students' listening ability did not show a sufficient range in terms of difficulty. That is, there were no items that could measure high-level students' listening ability. The items are mostly concentrated at the middle level. Therefore, it seems that the tests do not function well enough to measure the listening ability of this group of students. However, the separation reliability coefficient was greater than 0.50; thus, the examinee ability measures were not mainly due to measurement error (Fisher, 1992). In addition, all examinees' infit mean-square values were within the acceptable range of 0.7 and 1.3 for a multiple-choice test (Bond & Fox, 2015), with only one exception of an infit mean square value of 1.4, which suggests that the majority of students' response patterns functioned as the model expected. That is, the estimated ability measures were a valid predictor of students' listening ability.

The item facet was also examined. As found in Figure 2, items 3 and 7 (1.29 and 1.16 logits, respectively) were most difficult, while item 5 (-2.88 logit) was the easiest, showing a 4.17 logit spread. The most difficult two items and the easiest item were all detail related. The two most difficult items required students to identify information that is inconsistent with a dialogue/monologue, while the easiest item asked students to identify what the speaker asked her interlocutor to do. That is, students might have felt cognitively burdened when paying attention to all of the details in order to find inconsistent information in the five options (items 3 and 7), compared to paying selective attention to specific information (item 5). Thus, it seems that it is not the target skill (finding details) itself that makes a

difference in item difficulty, but rather the amount of information to process that contributes to item difficulty. The strata value of 5.08 suggests that 17 midterm test items were roughly separated into five statistically distinct levels of difficulty. The separation reliability coefficient (R = 0.93) supported this distinction. That is, the items were well distinguished in terms of difficulty because the true variance far exceeded the error variance (Myford & Wolfe, 2000). The infit mean-square indices ranged from 0.83 to 1.30, all being within the acceptable range of 0.7 and 1.3, which implies unidimensionality. In other words, all 17 items on the midterm test contributed to the measure of a single trait (i.e., listening ability) (Bond & Fox, 2015; Eckes, 2015). Overall, the midterm test adequately measured students' listening ability, while there was a lack of more difficult items.

The analysis of the final test performance displayed a very similar pattern to the results of the midterm test analysis (for this reason, the all-facet map for the final test is not presented here). The examinee facet showed a 6.82 logit spread (from -2.39 to 4.43), and the strata value was 2.40 (R = 0.71), suggesting that students' listening ability could be separated into approximately three levels. However, separation without a very high reliability coefficient implies that the examinee ability estimates are somewhat due to error variance. The same reason, as mentioned in the midterm test analysis (weak alignment between students' listening ability levels and item difficulty levels), might have attributed to this error variance. Similar to the midterm test, all examinees' infit mean-square values were within the acceptable range of 0.7 and 1.3, with only two exceptions. Thus, overall the final test seems to be valid in predicting students' listening ability.

The final test items showed a 4.44 logit spread in difficulty (-2.71 to 1.73). The items were separated into approximately five statistically distinct difficulty levels (strata value of 5.41, R = 0.94). The high separation reliability estimate supports the strata value. In addition, all 17 items fit the model, thus supporting the validity of the test items. Very similar to the results from the analysis of the midterm test performance, the analysis of the final test provided evidence of reliability and validity of the test results.

Conclusion

This study examined the validity of classroom assessments used in an actual EFL classroom (midterm and final tests). The classroom-based listening tests covered the curricular learning goals and target skills practiced in the instruction. However, there was an imbalanced use of learning goals and target skills for these tests. The tests also differentiated among the students in terms of their listening ability, but failed to provide evidence of a range of item difficulty that could contribute to a better estimation of students' listening ability. Therefore, it seems there is substantial room for improving the validity of classroom assessments. Improving the validity of classroom-based assessments is important in order to gain information from these assessments, which can trigger subsequent L2 teaching and learning. In other words, assessment *for* learning can be anticipated beyond assessment *of* learning.

The current study showed how classroom assessment can be validated with regard to a specific learning context (including curriculum, instruction, and assessment), which gives insights into the development and use of L2 classroom assessment. The study revealed how important it is to consider the curriculum and instruction for the development of L2 classroom assessment in order to assess what has been targeted and taught in the classroom. In order to help L2 teachers develop and use classroom assessment appropriately, as discussed in the validity analysis framework, continuous support and collaboration from school districts, teacher educators, and L2 teachers themselves are key to teachers' successful implementation of classroom assessment. The current study also applied the validity analysis framework developed for instructional settings, suggesting that different components of validity need to be considered for validating L2 classroom assessment due to its unique nature. Validity argument frameworks proposed mostly for large-scale standardized tests might not sufficiently evaluate the validity of classroom assessment.

Despite its implications, the current study provides a single example regarding the validation of L2

classroom assessment. For future research, it may be worth investigating how the teacher has developed tests in relation to the curriculum and instruction, and how students react to these tests. Doing so will bring about a better understanding of L2 classroom-based assessment because the teacher and students are the central components of LOA. Learning from assessment cannot be expected without the teacher or students' active engagement. Also, other language skills included in the curriculum, such as reading and speaking, should be considered together with listening skills to evaluate the overall validity of classroom assessment.

Acknowledgements

This work was supported by Hankuk University of Foreign Studies Research Fund.

The Authors

Hyun Jung Kim is an associate professor of the Graduate School of TESOL at Hankuk University of Foreign Studies in Seoul, Korea. Her research interests include second and foreign language assessment, learning-oriented assessment, and language test validation. Her recent publications include articles on validation of L2 performance assessments, effectiveness of instructor feedback for learning-oriented language assessment, and an analysis of rater behavior on an L2 speaking assessment.

Graduate School of TESOL
Hankuk University of Foreign Studies
107, Imun-ro, Dongdaemun-gu, Seoul, 02450, Korea
Tel: +82-2-2173-3154
Email: hkim@hufs.ac.kr

Sung Su Yang is an English teacher at Sung-ho High School in Gyeonggi-do, Korea. She earned an M.A. in English Education from the Graduate School of Education at Sungkyunkwan University in Seoul.

Sung-ho High School
46, Dongbu-daero 436beon-gil, Osan-si, Gyeonggi-do, 18150, Korea
Tel: +82-31-371-7161
Email: excelsior11@naver.com

References

Ahn, B-K., Oh, Y-J., Kim, A-S., Choi, H-J., & Kim, A. D. (2013). *High school English conversation.* Seoul, Korea: Chunjae Education Inc.

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

Bachman, L. F. (2005). Building and supporting a case for test use. *Language Assessment Quarterly, 2*, 1-34.

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

Bond, T. G., & Fox, C. M. (2015). *Applying the Rasch model* (3rd ed.). New York: Routledge.

Brindley, G. (2001). Outcomes-based assessment in practice: Some examples and emerging insights. *Language Testing, 18*, 393-407.

Brown, J. D. (2005). *Testing in language programs: A comprehensive guide to English language assessment*. New York: McGraw-Hill.

Butler, Y. G., & Lee, J. (2010). The effects of self-assessment among young learners of English. *Language Testing, 27*, 5-31.

Cheng, L., Rogers, T., & Hu, H. (2004). ESL/EFL instructors' classroom assessment practice: Purposes, methods, and procedures. *Language Testing, 21*, 360-389.

Chon, Y. V. (2014). Lexical threshold of L2 reading in the Korean CSAT. *Journal of British and American Studies, 31*, 341-376.

Colby-Kelly, C. (2014, October). *A theoretical analysis approach to AFL pedagogical materials development in an L2 classroom setting*. Teachers College Columbia University Roundtable in Second Language Studies, New York, U.S.A.

Colby-Kelly, C., & Turner, C. E. (2007). AFL research in the L2 classroom and evidence of usefulness: Taking formative assessment to the next level. *Canadian Modern Language Review, 64*, 9-38.

Eckes, T. (2015). *Introduction to many-facet Rasch measurement: Analyzing and evaluating rater-mediated assessments* (2nd ed.). Frankfurt, Germany: Peter Lang.

Educational Broadcasting System. (2017). *Special lecture for College Scholastic Ability Test: English listening*. Seoul, Korea: Educational Broadcasting System.

Fisher, W. P. Jr. (1992). Reliability statistics. *Rasch Measurement: Transactions of the Rasch Measurement SIG, 6*, 238.

Green, T., & Hamp-Lyon, L. (2015, March). *Learning oriented language assessment for the classroom: A primer*. Workshop presented at the 37th Language Testing Research Colloquium, Toronto, Canada.

Hughes, A. (2003). *Testing for language teachers* (2nd ed.). Cambridge: Cambridge University Press.

Kane, M. (2006). Validation. In R. Brennen (Ed.), *Educational measurement* (4th ed.) (pp. 17-64). Westport, CT: Greenwood Publishing.

Kane, M. (2012). Validating score interpretations and uses. *Language Testing, 29*, 3-17.

Kim, A-Y., & Kim, H. J. (2017). The effectiveness of instructor feedback for learning-oriented language assessment: Using an integrated reading-to-write task for English for academic purposes. *Assessing Writing, 32*, 57-71.

Korea Institute for Curriculum and Evaluation. (2012). *Research on the development of achievement goals and levels for the revised English national curriculum*. Seoul, Korea: Korea Institute for Curriculum and Evaluation.

Kwon, S. K., Lee, M. B., & Shin, D. K. (2017). Educational assessment in the Republic of Korea: Lights and shadows of high-stake exam-based education system. *Assessment in Education: Principles, Policy & Practice, 24*, 60-77.

Leung, C. (2004). Developing formative teacher assessment: Knowledge, practice and change. *Language Assessment Quarterly, 1*, 19-41.

Leung, C., & Lewkowicz, J. (2006). Expanding horizons and unresolved conundrums: Language testing and assessment. *TESOL Quarterly, 40*, 211-234.

Leung, C., & Scott, C. (2009). Formative assessment in language education policies: Emerging lessons from Wales and Scotland. *Annual Review of Applied Linguistics, 29*, 64-79.

Ministry of Education, Science, and Technology. (2009). *English National Curriculum*. Seoul, Korea: Ministry of Education, Science, and Technology.

Myford, C. M., & Wolfe, E. W. (2000). *Monitoring sources of variability within the Test of Spoken English Assessment System*. Research Report No. 65. Princeton, NJ: Educational Testing Service.

Pellegrino, J. W., DiBello, L. V., & Goldman, S. R. (2016). A framework for conceptualizing and evaluating the validity of instructionally relevant assessments. *Educational Psychologist, 51*(1), 59-81.

Purpura, J. E. (2016). Second and foreign language assessment. *Modern Language Journal, 100*, 190-208.

Rea-Dickins, P. (2006). Currents and eddies in the discourse assessment: A learning-focused interpretation. *International Journal of Applied Linguistics, 16*, 163-188.

Rea-Dickins, P., & Gardner, S. (2000). Snares or silver bullets: Disentangling the construct of formative assessment. *Language Testing, 17*, 215-244.

Tsagari, D. (2014, October). *Unplanned LOA in EFL classrooms: Findings from an empirical study*. Teachers College Columbia University Roundtable in Second Language Studies, New York, U.S.A.

Turner, E. (2012). Classroom assessment. In G. Fulcher & F. Davidson (Eds.), *Routledge handbook of language testing* (pp. 65-78). New York: Routledge, Taylor & Francis Group.

Turner, C. E., & Purpura, J. E. (2016). Learning-oriented assessment in second and foreign language classrooms. In D. Tsagari & J. Banerjee (Eds.), *Handbook of second language assessment* (pp. 255-273). Berlin, Germany: De Gruyter Mouton.

Yin, M. (2010). Understanding classroom language assessment through teacher thinking research. *Language Assessment Quarterly, 7*, 175-194.