# Balancing Validity, Practicality and Reliability on an University Oral English Test

Carlos Ramirez

## Abstract

Spoken English is becoming a topic of growing concern among Japanese government policy makers. The main question is how to achieve fluency given that past methodologies have not achieved their expected results. Testing as a tool for learning can help in attaining proficiency. However, a good test must adhere to the three guiding principles of validity, practicality and reliability. This paper proposes one type of testing, the Unified Oral English Test, as an acceptable method of testing at the large institutional level, i.e., at universities. It is proffered that this test conforms to the principles of validity and practicality. While it is acknowledged that confirmation of the test's reliability requires further data, the test content and organization suggests reasonable reliability. The first section of the paper explains the importance of test construct and content in test design. This explanation is followed by a general outline of the test itself. The remaining parts of the paper are discussions of the four main types of validity as they relate to the test, the practicality of the test and finally the challenges of achieving reliability on a Unified Oral English Test.

**Key Words:** Unified oral testing, validity, practicality, reliability, objectivity and washback

## Introduction

The wheels have been slow in turning but the prioritization and direction of English language studies within Japanese society have gradually and finally become clear. There has been a distinct and definitive movement towards an interest in oral competence and communicative language.[1] In the past, lip service was paid to English oral competence, yet not much was done in terms of concrete action. However, in the past few years this has changed. This change can be seen at the elementary school level with the introduction of English conversation classes. In Japanese society at large, the rising interest in spoken competency can be seen in the introduction of an oral component in the TOEIC test and a renewed interest in the TOEFL test, particularly the iBT version with its oral component. At the university level, the growing number of English oral classes taught by both native and non-native instructors also demonstrates the growing importance of the speaking aspect in English language education. The current Japanese government has made it clear that English language education will be an important and essential plank in an overall drive to reform and reinvigorate the economy (ICEF Monitor, 2013).

This trend notwithstanding, at least at the university level, there is a worrisome lack of instruments to measure actual progress in oral competence. This has contributed to unsatisfactory results in the competency of graduating students. To be sure, there also needs to be a focus on the curriculum and syllabi of courses in order to strengthen university level English programs. Faculty development should also be on the agenda to ensure that well-trained, competent teachers are in the classroom. In terms of testing, however, first and foremost it should be recognized that evaluation is just one of a number of integral parts of an English program. That is, testing should not only be used as a tool to measure progress in competency but also as an instrument of language acquisition. One of the ways to accomplish this is through positive washback. By being aware that an oral test will be conducted at certain intervals of the course, students will prepare by making efficient and effective use of in-class time during oral activities. As Kitao and Kitao (1996) point out, "Testing speech is important for its washback effect, even if the methods of testing and of assessment are not as perfect as they might be" (p.7). Despite the imperfections of oral testing, as noted by Kitao and Kitao and others, most researchers and classroom teachers find much value in testing as a means to language acquisition.

It is not uncommon for oral competency at universities in Japan and other foreign countries to be measured haphazardly due to the complicated logistics of testing. Japanese university English classes are characterized by large numbers of students, limited time and insufficient resources (see Gong, 2004, for Taiwan). Even if these issues can be resolved to satisfaction, speaking tests at universities suffer from a number of other flaws relating to their validity, practicality and reliability. Given the general sense of urgency in Japanese society to improve oral communication, this author and his colleagues have implemented an oral test as part of the English language curriculum for non-English majors. It is called the Unified Oral Testing format.

The principles of validity, practicality and reliability are universally acknowledged as basic guidelines for creating an effective test (e.g., Al-Amri, 2010). This paper argues that a carefully designed test can for the most part conform to these basic principles despite the numerous theoretical and practical challenges administrators face at the university level. Specifically, the author will begin by noting and discussing how close attention to the importance of test construct and test content can enlighten administrators to the main issues of test design. This discussion will be followed by a description and general outline of test format created by the author and his colleagues. The paper will then proceed to examine the faculty Unified Oral Testing format in light of the four main types of validity: construct, content, predictive and face validity. In regards to practicality, this paper will show that by being creative with administrative logistics, it is possible to achieve testing objectivity, i.e., fair and impartial assessment of students. Sound test validity and objectivity has a profound washback effect on students. Finally, the reality of limited financial resources will demonstrate how budgetary issues have a negative impact on reliability.

## Test Construct and Content: What are we testing?

The first step in developing any oral English program curriculum is to define the construct of speaking. What does speaking English entail? There is much literature and debate devoted to this topic and this paper does not have the space or scope to engage in it (for good discussions see Fulcher, 2003; Luoma, 2004; and, especially, Bachman 1990). The main controversy centers on whether language is contextual, dependent on the interaction between two people, or is it independent of context and fixed even before any interaction takes place (Fulcher, 2003). Suffice it to say that most researchers and practitioners in the field have opted for a hybrid theory. As Bachman (1990) notes, speaking entails having knowledge of, or competence in, the language, and the ability to execute, or produce, that knowledge within a variety of contexts.[2] Beginning with Hymes (1972), this hybrid approach has evolved into the Communicative Language Ability approach to teaching and is now widely accepted within the field of ESL as the mainstream teaching approach to English. The main supposition of this approach is that language is primarily a tool for communication. Communication, according to this theory, requires quite a wide skill set that includes grammar, discourse and sociolinguistic competence as well as competence in functional language (Harsono, 2005; Miayata-Boddy & Langham, 2000; Bachman 1990).

Once an underlying construct for speaking is agreed upon, (be it the above mentioned or not) then program administrators can turn their attention to the task of test development. As stated earlier, one of the main objectives of a communicative speaking test is its contribution to overall language proficiency. Moreover, and beyond this fundamental goal, specifically a test must measure what it is supposed to be measuring. Put another way, the test content must reflect both a valid speaking construct (i.e. a valid definition of speaking) and the in-class materials that are being tested. This is called content validity. Within a university context, this means the test should match the content of the departmental curriculum and the class syllabus. These, in turn, should also be derived from a well-defined construct.

According to Munby (2004), a university test is valid if it measures the following two points:

1) It must measure mastery of course content.

And, more generally,

2) It must measure speaking skills and level of communicative competence in English.[3]

While these two goals seem rather innocuous and obvious, there is considerable debate as to what the content on the test and, hence the measurement, should reflect. Depending on the content, the test taker's output will also differ significantly and it may or may not match the original construct for speaking. In more concrete terms, if mastering course content is the main focus of the test content, then students will skew their study strategies to only reviewing the tasks learned in class. Thus, the students' final scores on the test may not represent their real language capability and the overall level of the student's proficiency. It may only reveal the student's ability to memorize and to repeat set patterns and expressions that were studied in class.

Alternatively, if the test is mainly measuring general language ability with test content that reflects this, then students are not likely to pay as much attention to specific tasks and structures taught in the classroom during the term. Students will tend to rely more upon their latent language abilities during testing time. Accuracy suffers the most when students follow this strategy and can often lead to unintelligible output. Therefore, emphasis on measuring a student's general language level could negate any potential positive washback effect on students' motivation to concentrate in class.

So, what is the test developer to do given this dichotomy regarding content? Drawing on Weir's work, Munby (2004) states that test content (he uses the term "operation") should involve both informational and interactional routines as well as improvisational skills. By informational and interactional routines, he is referring to materials, structures and tasks taught in the classroom. Improvisational skills, on the other hand, are general language skills that "involve negotiation of meaning and management of interaction" (Munby 2004, p. 136). Therefore, there need not be a conflict between measuring student mastery of class content materials and measuring general language proficiency as long as the test prompts students to generate language that includes both informational/interactional routines and improvisational skills.

In more general terms, Bachman (2002) describes this duality of purpose in testing as the distinction between a "task-centered" approach and "construct-centered" approach to testing. The task-centered approach focuses on authentic, concrete tasks that correspond to tasks used outside the test itself and engages test-takers in authentic language (Bachman, 2002, p. 455). Conversely, the construct-centered approach focuses on general language ability as defined by "the knowledge and skills that underlie the language construct" (Chalhoub-Deville as quoted by Bachman, 2002, p. 455). However, like Weir and Munby (2004), Bachman notes that the approaches are not necessarily mutually exclusive. On the one hand, those proposing a task-centered approach recognize that the inferences we make about language output on these tests are associated with the underlying language ability and students' capacity for language use (Bachman, 2002). Brindley (cited in Bachman, 2002) explicitly defines task-based assessment as demonstrating both knowledge of and the ability to use the language. On the other hand, construct-based approaches emphasize that a well-designed test should measure underlying knowledge of the language but referenced through specific tasks that encompass the target language use domain. In sum, according to Bachman (2002) "sound procedures for the design, development, and use of language tests must incorporate both a specification of the assessment tasks to be included and definitions of abilities to be assessed" (p. 457). Luoma (2004) concurs and recommends that test developers clearly state in writing the test "construct specifications". These specifications should include the purpose of the test, an overview of the tasks and rating criteria and, finally, describe the speaking skills to be assessed in a clear and concrete way (Luoma, 2004, p. 116). In so doing, developers through the construct specifications should be able to make clear that the test ***can*** and ***will*** measure specific task based language within a context of general language capability.

Once test developers are cognizant of these two main purposes and approaches to testing and for the need to reach a synthesis, the following four questions come to mind: 1) What content and tasks will accurately reflect the construct and lead to positive washback? 2) What format should the test take to enhance objectivity and credibility? 3) How do we assess performance on the test? 4) How do we attain some satisfactory level of reliability given the constraints of context especially at the university level? These questions must be taken within the context of the dual goals and approaches as explained above. These questions will be answered in turn in the following sections while proposing options based on a test created by the author and colleagues teaching in a first year language program at the university level (henceforth called "the faculty test"). Question number three will only be touched upon lightly. Assessing performance, the main focus of question three, requires substantial and detailed discussion as it relates to assessment criteria and rubrics and will not be discussed here. These areas will be addressed on their own in a separate forthcoming paper. First, however, it is necessary to give a brief overview of the test format itself as created by the author and colleagues of the faculty.

## The Unified Oral Testing Format: The Faculty Test

### Context

The English language program began in April 2010 in conjunction with the founding of the faculty. The faculty teaches social sciences to undergraduate students. Hence, the English program is geared towards students whose major is not English. Approximately 500 students enroll each year into the first year program of the faculty. During the first year, the English program comprises of two core courses: Oral English 1 and 2 and Eigo Enshu 1 and 2. The Oral English 1 course uses a unified syllabus created by full time faculty. Instructors are asked to teach the skills and topics listed in the syllabus so that all students are exposed to the same content. The Oral English class is taught by a native English speaker. The Eigo Enshu class is taught by a Japanese teacher of English. Both classes meet twice a week with the same instructor. All students must enroll in both classes. Students are assigned to a class consisting of 24 to 30 students after taking a placement test. Classes are divided into eight to ten different levels. In general, student proficiency level upon entrance into the faculty English program is approximately A1 to A2 as measured by the Common European Framework of Reference or 300 to 380 points as measured by TOEIC. Oral Testing takes place twice a semester: there is a mid-term test and a final test. All first year classes take part in the testing at the same time on these two days of testing.

### Test Outline

There are two parts to the test, a pair conversation which takes place on day one of testing and a teacher-student interview which takes place on day two of testing. Teachers do not test their own students. They exchange classes with the teacher directly above or below their own class level.

### Pair Conversation Test.

The content of the test involves general role play scenarios in which pairs are given a topic and are expected to expand on the topic by asking and responding to questions with their partner.

#### *Procedure.*

Approximately two weeks prior to the test, students are given 3 to 5 topics based on tasks selected from the syllabus. Students are told that on testing day, they will be asked to perform one of the role plays but will not know which one since it will be selected randomly moments before their test. Students will not know their partner either until moments prior to the test as each pair of students is selected at random. On testing day, two students form Pair A and are given a topic from the list. Pair A has 3 to 4 minutes to rehearse. Before Pair A begins their test, a second group, Pair B, will be given one of the topics from the list. While Pair A is taking the test, Pair B is rehearsing at a designated "rehearsal" location within the classroom. When Pair A completes its test, then Pair B begins its test. While Pair B takes its test, two students have been selected to form Pair C and they begin to rehearse on a topic from the list and so on with Pair D, Pair E, etc.[4] Students are given approximately three minutes to complete their role play. Pairs are encouraged to keep talking until the teacher gives a finishing signal. At that point, students have 10 to 20 seconds to naturally close their conversation. Here are two topic examples from a mid-term test:

***Example 1.***

Topic/Task: Talking about daily activities.

Your conversation should include the following points:

1) A greeting; 2) Discuss what you usually do in the morning, afternoon and evening; 3) Ask for repetition; 4) A closing; 5) Other.

***Example 2.***

Topic/Task: Talking about your family.

Your conversation should include the following points:

1) A greeting; 2) Each partner should ask each other about their family; 3) Each partner should talk about their interests. What do they look like? What are they like? What do they do? 4) Ask for repetition or clarification; 5) A closing; 6) Other.

### Teacher-Student Interview Test.

This part of the test is titled Interview but it is better defined as a conversation between the teacher and student. In the instructions given to students two weeks prior to testing, it is clearly noted that the interview should be conducted as a conversation rather than an interview. Teachers are also briefed on creating a test atmosphere conducive towards a conversation-like format. It is recommended teachers interject when students begin to reproduce a long, memorized speech. Students are also encouraged to ask the teacher questions as well as to respond to questions from the teacher. The interview lasts approximately 3 minutes. The topic of the interview is based on the students' essay topic in their Eigo Enshu classes taught by a Japanese English speaker. An example of a topic is listed here:

***Example 1.***

Topic/Task: Your Old High School

Describe your school. Illustrate what your school is like, including its sights and sounds, the neighborhood or environment it is in, and anything about it that makes your school unique. It is OK to include memories of high school, for example, a school festival, a sports festival, a school trip, club activities or whatever you remember doing there. Through these episodes, the listener who knows nothing about your school should be able to "see" it in their mind.

## The Validity of the Faculty Unified Testing Format

There are four main types of validity in testing: construct, content, predictive and face validity. In this section, the author will examine the faculty test against each type of validity.

### Construct Validity in Theory and Practice on the Test

The curriculum of the faculty program is based on the Common European Reference for Languages (CEFR).[5] According to the Research and Validation Group (RVG) of the University of Cambridge (2009), CEFR itself is predicated on the models built by Canale and Swain (1980) which were then further developed by Bachman (1990) into the Communicative Language Ability approach. Like this approach, the faculty curriculum is grounded in the view that the ability to speak involves multiple competencies including grammatical knowledge, pragmatic awareness and phonology (RVG, 2009, pg. 4). In addition, the curriculum and syllabus of the core courses recognize the importance of the functional use of language: communication must be purposeful and goal-oriented within a specific context. Both Bachman's (1990, 2002) and CEFR's (RVG, 2009) approach to task-based learning emphasize the usefulness of this approach as a learning and assessment tool for functional and general language knowledge. As Luoma (2004) explains, tasks can be viewed as "language use situations, and making test developers compare test tasks with non-test language-use tasks may be very useful" (p. 101). Luoma (2004) further suggests that if the task is too limited, inferring that they may not replicate everyday language ability, then another task can be added.

CEFR advances task-based learning by simplifying tasks into Can-Do abilities, i.e., whether learners are able to perform certain activities using the target language. The latter lends itself to adaptation as a particular learning or institutional context may want to emphasize a different set of abilities or, in CEFR terms, Can-Do skills. Similarly to CEFR, Can-Do statements form the main foundation of the faculty curriculum. The Faculty Can-Do statements were, indeed, adapted from CEFR's lists but were rewritten to better match the low starting levels or abilities of incoming students. The four lowest

CEFR levels became the main focus of attention for the Faculty administrators and were divided into 5 levels as follows. As noted in Table 1, The K-Global level is the highest level and is equivalent to CEFR's B2 level but at an upper level. The lowest level is K-4 and equivalent to CEFR upper A1/lower.

**Table 1 CEFR and Faculty Levels**

| **CEFR Equivalency** | **Faculty Level** |
|---|---|
| Upper B2 | K-Global |
| Upper B1/Lower B2 | K-1 |
| Lower B1 | K-2 |
| Upper A2/Lower B1 | K-3 |
| Upper A1/Lower A2 | K-4 |

As mentioned earlier, most students fall into the K-4 category. To a large extent the testing assessment scales follow the description of the K-4 level for the core courses in the 1st year program. A forthcoming paper will discuss in depth the traits, skills and grading bands in the rubric used for the faculty test. For now, as an example, the overlap between CEFR and the Faculty can be seen in the evaluation of spoken production. The general description of speaking in terms of global self-assessment for both scales is as follows:

- CEFR A2/Spoken Production:

"I can use a series of phrases and sentences to describe in simple terms my family and other people, living conditions, my educational background and my present or most recent job (CEFR, 2001)."

- Faculty Can Do Description/K4 Level Descriptor - Speaking:

"I can use expressions frequently used in daily life and use set phrases used for social interactions naturally. I can speak for a half a minute about my family, people around me, living conditions, and school life."

For testing, the faculty rubric draws heavily from CEFR's analytical descriptors and task specific scales. For example, the descriptors for fluency and grammar are:

- CEFR A2/Fluency

Can make him/herself understood in very short utterances, even though pauses, false starts and reformulation are very evident (Council of Europe, 2001).

The fluency skill in the rubric of the Faculty Pair Conversation test is:

- Faculty K-4/Fluency/Pair Conversation Test

Speech tends to be slow with some hesitations but does not demand unreasonable patience of the listener. Speed of speech is reasonably paced albeit with a number of pauses.

- CEFR A2/Grammar

  Uses some simple structures correctly, but still systematically makes basic mistakes (Council of Europe, 2001).
- Faculty K4/Grammar/Pair Conversation Test

  Can produce basic tenses such as present, present continuous, past and future and, construct sentences with reasonable structure.

In general, all first year students are tested within the K-4 band. Within this band, the students are tested using a rubric comprising 5 criteria or skills rated one to five. Each skill is given a certain weight. Table 2 below shows the different skills and the weights for each skill. A total score of 30 or less would indicate a barely adequate grasp of the tasks being assessed and would mean the student is somewhere at the lower end of K-4 level (CEFR A1/A2). A grade of 40 or over would demonstrate an excellent performance of the skills and would suggest the student is towards the high end of the scale approaching K-3 (CEFR A2/A3) level.

The descriptor for each of the skills in the rubric reflect the course curriculum and rely considerably on CEFR's analytical and task scales as seen by the example descriptors provided above for fluency and grammar. The descriptor for each skill changes with the level: for example, Accuracy is described differently at each level from K-4 to K-Global.

Table 2 Pair Conversation Test and Point Score

| Criteria | Points |
|---|---|
| Fluency | 15 |
| Vocabulary and Expressions | 10 |
| Content | 10 |
| Grammar | 5 |
| Eye Contact and Body Language | 5 |
| Pronunciation | 5 |
| **Total** | 50 |

In sum, the level descriptors and testing assessment reflects the underlying speaking construct of the curriculum which in turn is founded upon models - i.e., CEFR and Bachman's Communicative Language Ability approach – that are well accepted within the field of applied linguistics. The criteria in the rubric represent fairly broad based skills used commonly in oral testing. They include both linguistics skills (fluency, vocabulary, accuracy and pronunciation) and task-based skills (content). This is not to say that the language models have been applied to perfection in the curriculum and on the test. However, it can be said that the ongoing research of both curriculum development and testing content is intent on finding common ground between underlying concepts and their effective application. Faculty administrators are keenly aware that the speaking construct, curriculum and test must adhere to common principles and concepts.

## Content Validity in Theory and Practice on the Test

The most basic tenet of test content validity is whether a test is measuring what it is supposed to be measuring. In essence, this means the test content must be a reflection of the construct (see previous section), and the content of the curriculum and syllabus. The Faculty Can-Do Handbook distributed to both teachers and students at the beginning of the year clearly identifies four main goals of the curriculum. Included in these goals are two aspects that are relevant to this paper. First, there is reference to the importance of developing discussion skills such as being able to give "one's own opinion" and to "interact with people...on their own initiative." Secondly, there is reference to improving abilities to work on "task". The Oral English syllabus for the 1st year core course is more specific as to what particular tasks are necessary to achieve language competency. The syllabus includes tasks or topics such as meeting new people, talking about daily activities, the future and the past, making and responding to requests and invitations among others.

In order to blend the curriculum goals and syllabus task topics, the test content has incorporated a fusion of the task-based approach and the construct based approach to testing as recommended by Bachman (1990) and Munby (2004). As emphasized by Munby, the optimum test for a course at the secondary or tertiary levels of education should measure both achievement of classroom tasks and a general improvement of language ability.

To attain the above goals of measuring classroom achievement, measuring general language proficiency and authenticity of task, the test is divided into two parts as noted earlier: a pair conversation among two students and a question and answer interview between a teacher and a student. I will discuss each part in turn.

### Discussion: Pair Conversation.

The objective of the pair conversation is two-fold: First, it is mainly an assessment of tasks as outlined by the topics in the syllabus. However, the role-play is designed as an open-ended discussion that will allow the partners to demonstrate skills beyond the narrow tasks of the syllabus. Second, it is hoped that the pair conversation will facilitate a more relaxed atmosphere for students in which they can demonstrate their competency and command of the language. It is also generally believed that pair conversations offer a less stressful situation as compared to the interview format since students form a more equal relationship when conversing with each other as opposed to with a teacher (see Taylor & Wigglesworth, 2009). It is also relatively easier for students in this format to mutually agree on a direction of conversation and to develop it in a more culturally and age-appropriate way. A stress-reduced atmosphere is further enhanced by granting a few minutes for rehearsal of the conversation prior to the actual test.

In both examples listed above, it can be clearly seen that the topic is a specific task – speaking about daily activities or talking about their family. However, the topic is broad enough for students to demonstrate their general latent knowledge and abilities of the language. At the same time, it gives them the opportunity to show their command of vocabulary and expressions taught during the semester. In eliciting both task-based and general language (i.e. construct-based language), teachers are able to gain speech samples that demonstrate competency of functional and discourse skills.

### Discussion: Interview Test.

The task of the interview test is very clear and the boundaries of the conversation are set. However, it is fair to conclude that the Interview is mainly using a construct-based approach to testing. The interaction and content of the conversation is fairly wide-ranging and extemporaneous.

One of the main objections to the Interview format is the dominant role of the assessor which makes it unlikely for test takers to take the initiative. In addition, it creates language patterns specific to "interview-talk" that are uncharacteristic of normal conversational speech (Fulcher, 2003). Given these objections, the interview format would seem to be in contradiction with one of the main goals of testing in that it should mirror real life speaking situations (Bachman, 1990). In general, this author would subscribe to these negative depictions of the interview test. However, one of the goals of the English program and a university education is to prepare students for life after university, which for most students means employment at a company. The Student handbook explains the purpose of English language learning as an experience to "enhance basic skills needed to become mature, responsible adults". This being the case, one can envision innumerous circumstances when students will be faced with challenging and stressful environments such as an employment interview; a situation requiring a verbal report, explanation or presentation to a superior; or an encounter after graduation in which the student must introduce him/herself to unknown people (even possibly in English). This interview test would, indeed, more than fulfill the requirement of mirroring real "stressful" life.

Because classes are twice a week, this English program has the luxury of time to test using both Pair Conversation and Interview formats. Not all English language programs can afford this amount of time. In the case there is insufficient time (i.e., insufficient number of classes), the merits of Pair Conversation would probably outweigh those of the Interview. Therefore, it would be recommendable to use the Pair Conversation format because it is a better reflection of everyday reality. The logistics of testing also favor pair conversations since it gives the assessor more quality time to evaluate the test takers as the assessor can focus attention on the students' performance of the task without having to participate in the task itself.

### Predictive Validity of the Test

The main purpose of testing is to assess a score on the ability of learners to perform some skill or set of skills. This score, in turn, must be an accurate indicator of that ability. It should also be a good predictor of how well the learner will be able to use the skill(s) in some future real life situation. This means the tasks performed on the test must also necessarily resemble to within reason the same task performed in real life. If not, then the score will not accurately measure the learner's ability to perform the task in real life. In other words, scores that do not accurately measure present or future performance are also worthless as indicators on which to base a judgment on employment or certification.

As previously mentioned, the evaluation criteria and scoring rubric of this test will be examined more closely in a future paper. To be sure, however, the rubric is constantly being reviewed by administrators and faculty. The rubric has undergone many changes since its inception and more research will be necessary to further its development. For now it suffices to say that at a minimum the test content and tasks as demonstrated by the given examples sufficiently resemble authentic interactions. Equally importantly, the test also measures what it is intended to measure, mainly, tasks taught during the term. With some confidence, it can be said that test scores given to students accurately reflect their abilities to perform, in the present and future, the tasks outlined on the test. These tasks, in turn, mirror everyday communicative interactions. Future research is necessary to confirm this hypothesis.

### Face Validity and Washback of the Test

Student and teacher feedback is a fundamental component to language learning. The positive or negative reactions of the students to classroom methodology, teaching content and testing will determine their overall progress. It is difficult to conceive of a situation in which students will attain high language competency within a negative, low motivated classroom experience. In reference to methodology and content, students need to be convinced that both will impact on their actual speaking skills. In reference to the test, it must act as a further incentive to acquire the skills taught in the classroom. The incentive within an academic institution such as a university is, of course, the final course grade. Proper testing content and implementation can affect student motivation by giving students a goal to focus on. As Gewirtz (1977) notes, there are few methods better than giving tests to achieve student competency of classroom taught materials. A test makes students concentrate more in the classroom and makes them more serious about their studies. However, testing must be implemented with all of the necessary trappings for students to perceive it as fair. That accomplished, the students will then accept their score as an objective assessment of their abilities (Gewirtz, 1977).

The unified testing format of the faculty thoroughly complies with Gewirtz's (1977) condition of proper formalities of test implementation. First, testing is done during the official mid-term and final examination periods of the school year giving it the due approbation of students. Second, because all students in the first year program are testing simultaneously, this creates an atmosphere of equal anticipation and anxiety. In all hallways and classrooms days prior to the test, one can often hear conversation among students pertaining to test preparations and content. Finally, as already mentioned, teacher exchange of classes gives an added sense of formality on testing day as students treat their unknown tester with appropriate decorum. In Japan, this is even more so the case as the word sensei (teacher) carries with it significant meaning in terms of deference from the speaker using it.

In general, the importance of face validity and washback cannot be underestimated. During the first four years of implementation of unified oral testing, students have enjoyed their oral classes as evidenced by class evaluation statistics: over 70% of students have rated the class as either motivating or very motivating.[6] Furthermore, the simple fact that the test is an oral one has had tremendous impact on student motivation and attitude towards speaking in class. Once students are aware that they will need to speak on the mid-term and final test, they intuitively engage in classroom speaking activities with an added sense of urgency.

### Matching Practicality and Objectivity on the Faculty Unified Test

Theory is a beautiful thing except when it bumps up with reality. Test administrators are always working within physical and financial restrictions. The key is to forge options that move testing closer to the ideal of theory. A unified oral testing format can be a realistic alternative to other forms of more efficient yet less valid testing systems. Most large Oral English programs at the university level opt for a written test for reason of convenience or efficiency. Faced with large a number of students and numerous classes, administrators find it too daunting to conduct a one-to-one oral test with each student. At most, teachers will conduct interviews independently of other teachers because they are too busy to co-operate with each other to workout test content and reliable assessment scales.

So, what are large institutional administrators to do given this situation? The unified testing format as described in this paper can overcome some of these practical issues by relying on program uniformity. This uniformity should be implemented by administrators. Program uniformity is defined as a uniform syllabus. That is, all instructors of the program teach the same fundamental skills, functions and concepts. By standardizing the syllabus, the testing content also becomes uniform. Administrators then only need to create one test and one set of assessment criteria for all classes thereby relieving individual instructors of this tedious duty. Of course, as implemented in this program, instructors are constantly being surveyed about the syllabus content, testing content and testing criteria to ensure that all stakeholders are in agreement with the underlying validity of the class and test content. The latter is crucial as a "stray" instructor teaching outside the parameters of the syllabus will put his students at a disadvantage when tested by another teacher. Similarly, the stray teacher will not be testing with the same mental conception of the assessment criteria. This will put the students this teacher is testing at a disadvantage.

The unified testing format can also match the principles of testing theory in terms of objectivity. A high standard of objectivity and fairness is achieved as all students are subject to the same test content and evaluation criteria. Objectivity is further enhanced by the fact that students are being tested by an instructor other than their own. Teacher exchange eliminates the teacher-student emotional attachment as students are unknown to the new teacher. It also means teachers will have no mis- or preconceptions of the students. Instructors are thus forced to rely on objective assessment scales for scoring rather than any personal bias in favor or against the students under examination. Furthermore, teachers are assigned to test classes at the level immediately above or below their own class. This results in less skewed scores as teachers are not influenced by the score of the first testee as the teacher is already familiar with the level.

## Reliability – The weak link of the Faculty Unified Test?

Reliability is one of the most fundamental aspects of testing. An unreliable test renders the scores meaningless in terms of their predictive validity. That is, a test score should measure how well a test taker will perform using the same skills from the test in a real world situation. However, if the scores are deemed unreliable, then there is no way of determining real world performance. In addition, test scores should be reliable so that they reflect test takers true command of the task and not just random error (Wells & Wollack, 2003). Furthermore, if scoring is inconsistent, it is not possible to know whether the score accurately measures the task at hand (Wells & Wollack, 2003). There are two main types of reliability in testing. First, there is intra-rater reliability, which concerns the ability of a tester to give consistent scores to test takers. Second, there is inter-rater reliability. This is the consistency of scores among different testers.

Both intra and inter rater reliability can be influenced by two phenomena. The halo effect is one type of abnormality in which the evaluator begins to generalize the score achieved on one component of the test to another unrelated component. For example, if a tester scored well in the pronunciation component of the test, the evaluator may be influenced by this to inflate scores on other components, say, grammatical skills even though it may not be warranted. Similarly, sequencing is another phenomenon that impacts on evaluators. Sequencing occurs when the evaluator applies the grading criteria inconsistently. This can occur as a result of fatigue. If the tester does not take a break during long periods of testing, his or her ability may become impaired. Sequencing can also occur when a weak or strong performance by one student influences negatively or positively on the next student. For example, an evaluator may score a mediocre student with high grades on a test simply because that test taker followed a number of very weak performances.

In order to neutralize these challenges in testing, evaluators need to undergo constant training. Trainers need regular exposure to and manipulation of testing scales, testing content and testing formats. Barnwell (cited in Fulcher, 2003) called this familiarization of the testing system "testing socialization". Barnwell found in his research that reliability coefficients were much higher in groups of evaluators who completed numerous and thorough test training sessions. Unfortunately, at the university level, financial resources are limited and this kind of training is accorded low priority. Moreover, most universities employ part-time instructors who do not have the time to participate in training programs on a volunteer basis. Limited resources have other consequences. It is not uncommon for classes to have more than 30 students per class, which restricts time allowed for testing even in the pair conversation formats. As a consequence, the language sample may not be large enough to make a fair judgment. In the end, without expanded budgets for training, reliability will continue to be a challenge at the university level.

Further statistical research in this particular faculty will be necessary to clarify to what extent reliability is indeed an issue. There are a number of ways a reliability study can be conducted. The most common method entails a number of assessors testing the same students. The results from different assessors are collated and, using correlation coefficients, it is then determined whether there is a high or low reliability (i.e., consistency) among the assessors. Another less involved method would be the gathering of test grading data from all the teachers and to examine score consistency among each other. This author hopes to conduct such research in the near future. For now, given the testing method of teachers testing classes directly above or below their level, test content, the experience of teachers, the straight forward descriptors, the standard rubric and, finally, the unified curriculum, it can be said that reliability may not be such a large issue.

## Conclusion

The current emphasis of government policy on English education, particularly the spoken component of the language, is clear. The question is how to achieve results. Older testing formats such as written and/or listening tests to assess speaking ability are counterproductive to this goal. At the university level, administrators are faced with the challenge of teaching and improving the oral language abilities of a large number of students. This paper has argued that testing must play a fundamental role in overcoming this challenge. Testing must be seen as one of a number of means to the end of oral proficiency. Indeed, the unified oral testing format as described here demonstrates that it is one plausible alternative to furthering the objective of oral language competency. Any testing system must consider validity, practicality and reliability as governing principles to a good test. The unified testing format as practiced by this faculty and explained in this paper reasonably conforms to the principles of validity and practicality of testing. While the content and organization of the test is supportive of reasonable reliability, further research and data collection must be done to establish reliability as fact.

On the four key types of validity, this faculty's test was shown to strongly adhere to the criteria of each type of validity. First, the construct validity of the test is based on the underlying concepts of the Common European Framework which in turn is grounded on the general principles of Communicative Language Ability developed by Bachman and others. For these researchers, speaking involves both competency and performance. That is, speaking is both having knowledge of a language and the ability to orally produce that knowledge within a particular context. Second, the content of the test conforms to the content of the curriculum, syllabus and classroom materials. The questions on the test take a task-based format as proscribed by Bachman (2002) and CEFR. Task based questions make it relatively easy to assess classroom achievement. In addition to eliciting classroom achievement language, this test also attempts to assess overall language proficiency by creating open ended tasks compelling students to manipulate and demonstrate language beyond the task. This is done through the use of a dual format of pair conversations and one-to-one interviews. Third, whether the test accurately assesses ability to perform in real life is a question that will need further research. A forthcoming paper on the rubric, assessment criteria and testing scales will discuss this issue in more depth. Suffice it to say here that the demonstrated validity of the test construct and content suggests that there may also be predictive validity to the test. Yet, it is acknowledged that more research is necessary in this area to confirm this claim. Finally, and perhaps most importantly, there is significant face validity to the test. The washback effect can be enormous if implemented properly. Students will respond to valid oral testing by preparing for the test and by fully participating in-class oral activities.

The unified oral testing format is both a realistic and practical alternative to testing at a university level. A unified syllabus will simplify the preparations for the test. Administrators can extract concepts and task from the syllabus and need only to prepare one set of questions and assessment criteria for the entire program. The main positive ancillary effect is objectivity and fairness in testing and assessment. Teachers test without any preconceptions of the student as the student is unknown to them. Class exchange among teachers ensures this outcome.

Reliability, the final principle discussed, could be the Achilles heel to the format. University budgets limit the amount of training administration can offer to teachers. Because of large class sizes, students' testing time is limited. Also, without sufficient and ongoing training, teachers will not become comfortable with the testing criteria and concepts. This can result in skewed scoring and mistaken feedback to the student. Financial constraints on English language education must be addressed adequately and honestly. Notwithstanding this caveat, the unified oral testing format is an option certainly worthy of consideration by other large educational institutions. It is an educational tool that can enhance the whole English language learning experience.

*The author would like to thank James Jensen for proof reading an earlier draft and for providing incisive comments on the substance of the paper. The author also expresses his gratitude to the Journal referees for lengthy and instructive comments on both the substance and formatting of the paper. Any remaining errors are the responsibility of the author.*

## Endnotes

1 For a good summary of the new emphasis on speaking and communicative studies as promoted by the Ministry of Education, Culture, Sports, Science and Technology see Stewart (2009). Stewart also reviews some of the past travails of English language education policy

2 See Miyata-Boddy and Langham (2000) for a general review of the historical origins and issues surrounding the competence vs performance debate.

3 Munby's (2004) goals for testing relate to his specific test. He states another goal, to complete or negotiate a task, but this is specific to his test and not to general testing which is the point I am trying to make.

4 Allowing students to practice in a rehearsal area for a few minutes immediately prior to their test was dubbed the "Batter-box" technique. This technique was explained to the author by John Carle, a colleague in the faculty.

5 For a detailed discussion of the links between our faculty curriculum and CEFR see Shimo and Nitta (2011).

6 This figure is based on Class Evaluation surveys done in July 2010 and July 2013. In each survey the figure was above 70%.

## References

Al-Amri, M.N. (2010). Direct spoken English testing is still a challenge worth bothering about. *English Language Testing*, 3 (1), 113-117.

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

Bachman, L.F. (2002). Some reflections on task-based language performance. *Language Testing*, 19 (4), 453-476.

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

Council of Europe (2001). *Common European Framework of Reference for Languages: Learning, teaching, assessment*. Cambridge: Cambridge University Press.

Fulcher, G. (2003). *Testing second language speaking*. Harlow: Pearson Longman.

Gewirtz, A. (1997). Some observations on testing and motivation. *ELT Journal*, 31 (3), 240-244.

Gong, B. (2004, November 12-14). *Considerations of conducting spoken English tests for advanced students*. Paper presented at the Thirteenth International Symposium on English Teaching, Taipei, Taiwan. Retrieved from http://www.scu.edu.tw/english/teachers/byron_gong/Considerations%20of%20conducting%20spoken%20English%20tests%20for%20advanced%20students.pdf

Harsono, Y.M. (2005). Developing communicative language tests for senior high school. *TEFLIN Journal*, 16 (2), 237-255.

Hymes, D. (1972). On communicative competence. In A. Duranti (Ed.), *Linguistic Anthropology: A reader* (pp. 53-73). Malden, Massachusetts: Blackwell Publishers, Ltd.

ICEF Monitor (2013, May 15). *Japan's ambitious proposals for higher education and language sectors*. Retrieved from http://monitor.icef.com/2013/05/japans-ambitious-proposals-for-higher-education-and-language-sectors/

Kitao, S. K., & Kitao, K. (1996). *Testing speaking*. Retrieved from ERIC Document Reproduction Service No. ED 398 261, 1-7. http://files.eric.ed.gov/fulltext/ED398261.pdf

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

Miyata-Boddy, N., & Langham, C. (2000). Communicative language testing – an attainable goal? *Bulletin of Tokyo Kasei Gakuin Tsukuba Women's University*, 4, 75-82.

Munby, I. (2004). Some issues, options and recommendations in the testing of spoken interaction for students of oral and general English at universities in Japan, *Hokkai Gakuen University Studies in Culture* 28, 133-145.

Research and Validation Group, University of Cambridge (2009). *Examples of speaking performance at CEFR levels A2 to C2*. Retrieved from http://www.cambridgeenglish.org/images/22649-rv-examples-of-speaking-performance.pdf

Shimo, E., & Nitta, K. (2011). Developing Can-Do check lists as a self-evaluation tool for university-level English classes. *Kinki University Center for Liberal Arts and Foreign Language Education Journal* (*Foreign Language Edition*) *2* (1), pp. 225-245.

Stewart, T. (2009). Will the new English curriculum for 2013 work? *The Language Teacher*, 33 (11), 9-13.

Taylor, L. & Wigglesworth, G. (2009). Are two heads better than one? Pair work in L2 assessment contexts. *Language Testing*, 26 (3), 325-339

Wells, C. S., & Wollack, J. A. (2003). An instructor's guide to understanding test reliability. *Testing & Evaluation Services. University of Wisconsin.*