The World Language Center Speaking Assessment: A Tool for Measuring Learning Outcomes

Lary MacDonald*, Malcolm Daugherty*, Shin’ichi Hashimoto* and Greg Lindeman*

*Associate Professor
*Lecturer
World Language Center, Soka University

Keywords: Learning Outcome, Oral Proficiency, Accountability, University English Education

Abstract

Learning outcomes have gained greater importance in the context of increased accountability in higher education. In Japan, oral proficiency in English as a Foreign Language (EFL) programs is one learning outcome that is being measured. As a result, evaluative instruments and assessment tasks to measure oral proficiency have been developed, critiqued and revised. This paper focuses on speaking assessments conducted at the World Language Center (WLC), Soka University. The development, rationale and methodology of the speaking assessment are explained and data gathered from a limited sample are analyzed. Based on these results, revisions to the speaking assessment are provided.

1. Introduction

Educational institutions have been under increasing pressure by governments to demonstrate efficiency and cost-effectiveness (Brindley, 2001). Accountability, simply defined as “answerability for performance” (Romzek, 2000, p. 22) has become the goal of governments as a means to monitor and assess the effectiveness of higher education institutions. As a result, learning outcomes, often referred to as outcomes-based assessment, standards, or benchmarks, have been forwarded as a means to confirm accountability and have gained growing recognition in higher education worldwide (Brindley, 1998). In the U.S., the Council for Higher Education Accreditation (CHEA) defines learning outcomes as “knowledge, skills, and abilities that a student has attained at the end (or as a result) of his or her engagement in a particular set of higher education experiences” (CHEA, 2003, p. 5). In Europe, The Bologna Process: Towards the European Higher Education Area has promoted learning outcomes as a common framework for higher education (Adams, 2004). The Association of
American Colleges and Universities has published a set of Essential Learning Outcomes which focus on intellectual and practical skills such as critical and creative thinking and quantitative literacy (2011). Learning outcomes for higher education have also been developed in the U.K. in areas such as Social Science, English, and Environmental Science (Melton, 1996).

As a result of this global movement towards increased accountability, higher education institutions in Japan are coming under greater pressure to meet international standards (Kushimoto, 2010). As a result, the Ministry of Education, Sports, Science and Technology (MEXT) has modified its policies to promote outcomes-based assessment by enacting a third-party evaluation system for all higher educational institutions in 2004. One of the main organizations conducting third-party evaluations is the National Institution for Academic Degrees (NIAD), an organization which is a strong proponent of an outcomes-based assessment (learning outcomes) approach (Kushimoto, 2010).

The Japanese government has also enacted policies in the context of English as a Foreign Language (EFL) instruction promoting learning outcomes. In recent policy documents concerning English language education, the following statement, “it is essential for them [children] to acquire communication skills in English as a common international language” reflects the government’s emphasis on oral proficiency (Ministry of Education, Culture, Sports, Science and Technology [MEXT] 2003, para. 4). In the context of university education, MEXT states “Each university should establish attainment targets from the viewpoint of fostering personnel who can use English in their work” (MEXT, 2003, para. 7). While the statement is vague and defers to each university to devise and implement their own specific English language curriculum, it does reflect a first step by MEXT to hold universities accountable for learning outcomes in the EFL curriculum.

However, devising, revising and assessing learning outcomes as defined by either a single institution or a governmental body is extremely complicated (Brown & Abeywickrama, 2010). As Nunan (2003) indicates through a multi-national exploration of English education in Asia, such factors as lack of access to education and the varying ability of individual teachers to skillfully deploy the latest instructional methods demonstrate the significant hurdles many countries must surmount in order to implement effective ESL / EFL programs.

Oral proficiency is one of many learning outcomes for which EFL university programs in Japan are held accountable. However, in fact, the value of oral proficiency assessments remains highly contested among linguists. The authenticity, validity and reliability of oral proficiency tests have been heavily critiqued (Bachman & Palmer, 1982; Lantolf & Frawley, 1985). In addition, oral proficiency tests have been criticized for potentially containing bias against Global Engishes (Jenkins, 2006; Taylor, 2006), giving inordinate attention to linguistic factors at the expense of diverse communication strategies (McNamara, 2006). Furthermore, researchers have commented on the impact of gatekeeping, a critique of the significant role that the standard global assessments such as TOEFL and IELTS have on the interpretation of oral proficiency (Raimes, 1990).

Several perennial issues need to be addressed before constructing an instrument for assessing oral proficiency. One chief concern is the distinction between holistic scoring which
is defined as “expressing an overall impression of an examinee’s ability in one score” (Luoma, 2004, p. 60-61) and analytic scoring, defined as “containing a number of criteria ... each of which has descriptors at the different levels of the scale ... and the examinees ... get a profile of scores” (Luoma, 2004, p. 68).

Research has elucidated several key findings that help clarify distinctions between holistic and analytic scoring in oral proficiency assessments. Halleck (1995) found significant correlation when results of analytic and holistic speaking assessments of the same three productive tasks (narrating, role play and questioning / interview) were compared. Xi (2007) corroborates this research in a study which found that analytic scores were highly resonant with holistic scores derived from the same data set. This high level of score resonance between holistic and analytic scoring appears reassuring, but in deciding which scoring method is most advantageous for a particular program, careful consideration must be taken concerning the benefits and drawbacks of each style. For example, while holistic scales are perhaps easier for evaluators to use, they do not provide detailed and individualized diagnostic data for each subject (Luoma, 2004). Analytic scales, on the other hand, offer a more detailed profile of oral proficiency, and provide raters rich guidance in decision making, although multiple categories for rating may cognitively tax raters (Council of Europe, 2011).

Another key concern in developing an instrument to measure oral proficiency is task validity. Evidence supports utilizing multiple tasks in speaking assessments. For example a study by Lee (2006) demonstrated that increasing the number of discreet tasks from one to five significantly maximized score reliability. The inclusion of multiple tasks and task variability also supports Brown and Hudson’s (1998) call for a diversity of tasks in speaking assessments. A diversity of tasks not only provides rich data for diagnosis, but also helps to avoid aberrations such as the “practice effect” (Halleck, 1995; Teng, 2007) which reflects not the subject’s oral proficiency but rather their ability to detect or intuit and then mimic patterns in the test structure itself. This orientation towards the inclusion of multiple sources of assessment however must be counterbalanced by the time constraints encountered in large assessment projects such as the current study. Indeed, multiple task types will perhaps provide a clearer picture of subject’s ability but can significantly lengthen the time necessary for assessment.

Other researchers have observed variability in score reliability and rater effect in interview-based assessments. Brown (2003), for instance, reminds us that while significant work has been done to study the discourse of individual interviews, little has been done to investigate the various raters and how their individualized approach to the interview can have dramatic effect on the rating process.

As is evident in this brief review of some prominent literature in the field, assessing oral proficiency remains a difficult undertaking. The speaking tasks employed, the type of scoring rubric (holistic / analytic) utilized, and the consistency of scoring among raters can significantly impact outcomes. However, as Huerta Marcias argues, the assessment of oral proficiency is relevant in EFL instruction due to its focus on what students can do with the language rather than what they can recall (as cited in Stroupe & Panda, 2007).

Bearing these considerations in mind, the
World Language Center (WLC), Soka University, embarked upon the task of developing a speaking assessment to measure student progress in oral proficiency. It was hoped that the results would provide evidence that the WLC was achieving its learning objectives, thus demonstrating accountability.

2. Development of the WLC Speaking Assessment Instrument

In 2002-03, a committee of five full-time faculty was appointed to develop an in-house speaking assessment instrument for the WLC, as part of an ongoing development plan to build accountability into WLC programs (Stroupe & Panda, 2007). The WLC Speaking Assessment was designed to measure student ability to produce and interact in English in a face-to-face situation. After investigating various tests available at that time, it was decided that an interview format would best serve departmental needs. This would allow for the spontaneity and freedom associated with real speech of both the interviewer and the interviewee. As several of those involved with the project were trained as IELTS examiners, a similar format was adopted.

In the process of constructing the speaking assessment, holistic and analytic scoring rubrics were considered. Because of potential weaknesses in holistic scoring (Brown, 2000; Pollitt & Murray, 1996; Taylor & Jones, 2001), and the advantage of analytic scoring which provides rich diagnostic data (Luoma, 2004), the WLC chose to develop an analytic scoring rubric. This rubric was divided into eleven bands, from 0 to 10. Students who received a score from 1-3 were ranked as novice students, 4-6 intermediate, 7-9 advanced and 10 superior. A score of 0 indicated that no evaluation was possible through lack of any communication or a “no-show” status for the interview.

Three components of student performance were to be judged: Fluency and Coherence, Vocabulary, and Grammar. The Fluency and Coherence component was used to assess the ability of the students to comprehend and respond fully and fluently during face-to-face communicative situations. The Vocabulary component was used to survey the ease of production and the extent of student lexical breadth to discuss a topic. The Grammar component was used to determine student ability to use correct and appropriate grammatical structures during communication. Descriptors for each component within each band were established as markers for the evaluator. For example, at Band 5, students are generally understood by a sympathetic interviewer; are able to discuss familiar topics and recycle vocabulary; and have good control of simple tenses and an emerging use of past tense.

Format of the Speaking Assessment

The format of the WLC Speaking Assessment was divided into three parts: the warm-up, the extended speech, and the formal interview. This format was similar to the IELTS Speaking Test format used from 2001 (IELTS, 2011) though abbreviated from the IELTS 11-to-15-min format to a length of 5 to 7 min so as to allow all the assessments to be completed within the time constraints at the beginning and end of each semester. Also, many of those involved with the development of the speaking assessment felt that the rigidity of an IETLS-style scripted format would impede the spontaneity of interaction between the interviewer and the student, so a less structured format with only a rudimentary list of sample topics or
questions for each section was developed to allow interviewers freedom during the assessment to change topics and create their own questions.

**The warm-up.**

The first section of the test, about 1 min in length, is a chance for the students to meet and greet the interviewer and acclimatize to an atmosphere of English through the exchange of basic information about themselves such as hometown, family, or other information about their daily life. While facilitating the free flow of information from the students, the interviewer should use this time to preliminarily judge student output regarding fluency and coherence, vocabulary, and grammar.

**The extended speech.**

This middle part of the test, about 1.5 to 2 min in length, is designed to give students an extended speaking opportunity. The topics of the questions allow the students to provide in-depth descriptions of people, events, vacations, jobs, interests, or current events. The students are informed of the question and allowed 30 s to prepare before answering. The students then answer the question by talking as long as possible, up to 1 min, without interruption from the interviewer. The interviewers may give verbal prods if the student stops before the time limit has expired. The examiner should scrutinize the student responses for flow, connected discourse, extended descriptions, vocabulary, and control of grammar.

**The formal interview.**

The final portion of the assessment, a 2.5-to-3.5-min segment, is designed as a formal interview in which interviewers ask questions requiring more complex discourse strategies such as comparison, support, justification, and speculation; more expansive vocabulary to-ward abstract and specialized topics; and more variety and complexity of grammar structures. Upon ending this section, the interviewer discharges the student from the assessment site and rates the language which the student produced using the band scale descriptors of the scoring rubric to determine the score.

**Scoring**

During the scoring process, each of the three components of the scoring rubric—(1) Fluency & Coherence, (2) Vocabulary and (3) Grammar—are scored independently. The final score is the average of the scores of the three components. It is also possible to give half points to bands 1 through 9 if some but not all of the criteria of the next higher band are met.

**Interviewers, Training, and Quality Control**

All full-time WLC faculty (FTF) served in the role of interviewer for the assessment. The small group involved in the development of the assessment instrument acted as trainers. Training of the interviewers in the format of the assessment instrument, and inter-rater reliability of the scoring rubric were completed in a series of professional development workshops. For the purposes of research and training, students were asked to sign permission forms allowing their assessment session to be videotaped. These videotaped sessions, informal observations, and informal interviews with the interviewers were used to evaluate the process and address problems dealing with execution and scoring.

**Implementation of the Test**

The first trials of the WLC Speaking Assessment were held in the fall of 2003 by the trainers to appraise the instrument, timing, and
manpower needed for full implementation in the spring of 2004. The overall scope of the speaking assessment was to evaluate all students—more than 1000 in total—who were taking WLC courses with a speaking/communication component at the beginning and end of the course. The speaking assessments were planned for the second week of the spring semester, after class rosters were finalized, in order to minimize the effect of instruction on student scores, especially for freshmen who had just entered university. As most courses offered by the WLC are year-long courses, a second test was administered during the penultimate week of classes in the fall semester.

Prior to the initial implementation of the WLC Speaking Assessment in the spring of 2004, all 13 FTF were trained and normed in the use of the assessment instrument. However, after the first round of assessments, it was realized that such a large number of student assessments was extremely burdensome for such a small group of interviewers, as they had to divert most of their class time and many of their non-teaching hours—often six to seven hours per day—during the week to conducting the assessments. In addition, finding assessment venues was also problematic.

The overburdening of the FTF prompted the addition of the part-time faculty (PTF) who taught WLC classes with a speaking/communication element into the process the following year. As the PTF had limited availability, their inclusion in the speaking assessment process was problematic. The training of PTF commenced with lunchtime and afterhours professional development workshops and the development of a training video. All PTF who taught WLC classes with a speaking/communication element as well as new FTF were trained and normed prior to the beginning of the 2005 assessment cycle. During the 2005 and 2006 cycles, PTF were only required to conduct speaking assessments during their regularly scheduled class times while FTF were required to conduct assessments during their regularly scheduled classes as well as in some of their non-teaching hours, often consisting of 4 to 5 hours per day during the assessment week.

Several of those involved in the initial development of the assessment served to continuously review the instrument and assessment process as part of an ongoing effort to confirm its validity. The assessment instrument was fine-tuned and minor revisions were implemented. Sample videotaped sessions were reviewed and problems noted, and individual interviewers were informed of format or scoring problems.

The WLC Speaking Assessment, as part of a departmental effort to construct an effective and reliable program, served several purposes: to be one of a series of standardized skills-based tests on which WLC courses were focused (others included writing, TOEFL, and TOEIC), to help develop a set of departmental benchmarks/learning outcomes by level and skill for WLC courses, and to build and maintain a database of the test data to help evaluate the effectiveness of courses and the test (Stroupe & Panda, 2007). The benchmarks/learning outcomes for speaking were adapted from the bandscale descriptors of the assessment instrument for the different levels of courses offered by the WLC. The data gathered in the database over the three-year cycle of speaking assessments showed evidence of moderate to significant improvement of speaking skills over the course of a year as well as
strong inter-rater reliability among those administering the assessment (Stroupe & Panda, 2007). Despite the usefulness of the data and the instrument, it was decided to suspend further assessment due to the enormous expenditure of time and effort and the stress placed on the faculty and facilities.

In the spring of 2011, the university administration asked the WLC to implement another cycle of speaking assessments as a tool for measuring learning outcomes within the university context. A small cohort of students was assessed for the purpose of determining whether the WLC Speaking Assessment could be reinstated as a valid instrument to measure learning outcomes. The results and implications of this 2011 cycle of speaking assessments are discussed in the remainder of this paper.

3. Methodology and Results

A total of 42 students participated in both the pre- and post-tests. Students were enrolled in two sections of an Upper-Intermediate level English for Academic Purposes class. Students at this level have TOEIC scores between 490 and 565. According to the Educational Testing Service (ETS), students with TOEIC scores in this range demonstrate an ability to infer the central idea and purpose of listening passages and written materials; can understand details of short and extended spoken exchanges; and can understand a broad range of vocabulary and grammatical structures in written text (Educational Testing Service [ETS], 2007).

Paired-samples t-tests were used to compare the pre- and post-speaking assessment results. A paired samples t-test is used when two mean scores are compared but the groups are not independent (Larson-Hall, 2010). This statistical tool can be used for measuring the same group at two different time periods using the same measurement instrument as was the case in this study. The statistical analysis is based on rater scores. Two of the raters had received formal speaking assessment training (IELTS), and two were trained in-house. Due to a shortened semester of 11 weeks rather than the normal 15 weeks, pre- and post-speaking assessments were administered in the 2nd and 9th week of classes. The results of the combined scores for the three assessment components (Fluency and Coherence, Vocabulary and Grammar), and each assessment component observed independently are reported in Table 1.

Table 1 reveals a statistically significant difference in the pre-test (M=3.9, SD=.78) and the

<table>
<thead>
<tr>
<th>Variable</th>
<th>Pre-test</th>
<th>Post-test</th>
<th>t (42)</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>M</td>
<td>SD</td>
<td>M</td>
<td>SD</td>
</tr>
<tr>
<td>Combined</td>
<td>4</td>
<td>.78</td>
<td>4.6</td>
<td>.73</td>
</tr>
<tr>
<td>Fluency and Coherence</td>
<td>4.1</td>
<td>.78</td>
<td>4.8</td>
<td>.79</td>
</tr>
<tr>
<td>Vocabulary</td>
<td>4</td>
<td>.85</td>
<td>4.48</td>
<td>.92</td>
</tr>
<tr>
<td>Grammar</td>
<td>3.88</td>
<td>.79</td>
<td>4.47</td>
<td>.76</td>
</tr>
</tbody>
</table>

Note. The combined variable is a collective variable including fluency and coherence, vocabulary, and grammar.

**p.<.01.
post-test (M=4.6, SD=.73) scores: t(42)=-5.92, p =0.000, for the combined variable. This suggests a statistically significant increase in students’ overall oral proficiency in the 8-week period between the pre- and post-tests.

In addition, paired-sample t-tests were also conducted separately for each bandscale criterion (Fluency and Coherence, Vocabulary, and Grammar) which also revealed statistically significant differences. Fluency and Coherence measures students’ ability to comprehend the question, respond with appropriate language, demonstrate their knowledge of the topic, and display various discourse strategies. Pre-test (M=4.1, SD=.78) and post-test (M=4.8, SD=.79) scores: t(42)=-5.77, p=0.000, reveal improvement.

To further examine the relative difference of each bandscale criterion between the pre- and the post-test, a small sample of student responses were transcribed and analyzed. Ten students signed consent forms agreeing to be videotaped for both the pre- and post-tests. Each video-taped sample was examined and 3 examples that were deemed most representative were selected for transcription. The bandscale categories of Fluency and Coherence, Vocabulary, and Grammar were analyzed. Additional objective criteria such as the number of words used and the length of responses were noted.

Excerpts 1 and 2 illustrate progress in the Fluency and Coherence component of the speaking assessment bandscales. Student A is responding to the prompt, “What kind of books do you like to read?”

Excerpt 1: Student A Pre-test

Mori Eto’s book is... umm is to take me to comfort and umm easy for me in reading them uunto (15 s pause) uunn I ah I like the books of Mori Eto because hers book is (laugh) umm is eto hers book eto. Ran is eto ah book is eto most like for me. Uh? un -un-un eto I like ah I most like eto because Ran is ah hero challenge long run marathon, and she...she is ah she practice very hard. I read when I’m high school student, and uh and to I courage this ah hero umm courage hero to unn (laugh) finish.

Excerpt 2: Student A Post-test

Oh...I like uunn heart-warming story uunn because I feel uun uummm ah when I read these books, I feel comfortable, and warm my heart uun I I ah these book encourage me when I ah when I in trouble. I when I junior high school, these books help me eto overcome these troubles uunto ah (laugh) especially I like books eto wro-written by Eto Mori. un Eto Mori. Eto Mori write interesting story and heart-warming story. Her her books eto eto her books change movie (laugh)... nanteiebai-indarou? make film, make films? But I don’t watch these movie, but I read these books. I want to watch these movie also.

The probability of students getting the same prompt during the extended speaking opportunity in the pre- and post-test is fairly low, but in this case, the student happened to receive similar prompts about books, which provided a rare opportunity to directly compare the responses. In general, during the pre-test, Student A’s responses were shorter and less developed than in the post-test. Speaking time during this part of the test increased from 2 min 35 s in the pre-test, to 3 min 58 s in the post-test, and the number of words used increased likewise from 128 to 198. In the first part of the response in the pre-test, there was a
lengthy pause when Student A tried to formulate the next utterance, whereas in the post-test, greater fluency was demonstrated. The presence of the L1, with words such as *eto* and *un*, was more prominent in the pre-test than the post-test, and it is interesting to note that Student A switched the order of family name and given name to fit Western conventions in the post-test.

Paired-samples t-tests were also performed on the Vocabulary component of the speaking assessment, which examines the extent and complexity of the lexis of students to discuss general / specific and concrete / abstract topics, and the ability to circumlocute and rephrase in appropriate situations. Pre-test (M=4. SD=.85) and post-test (M=4.8, SD=.91) scores; t (42)=-3.53, p=0.001, show student improvement.

In the pre-test, Student B used 111 discreet headwords out of 237 total words, whereas in the post-test, Student B used 141 discreet headwords out of 368 total words. In the post-test, expressions such as “It’s very delicious,” “You should see Nagoya castle,” “I am so negative,” “My mother always encouraged me,” and “Now I live by myself.” demonstrated more appropriate use of collocation. Also, Student B used circumlocution to rephrase an unfamiliar word. When trying to explain a famous local dish of “eel,” Student B described eel as a “long fish”. Overall, Student B was able to express ideas more effectively in the post-test, thus leading to a higher score.

Finally, the Grammar component reflected students’ ability to use correct and appropriate grammatical structures during communication. Pre-test (M=3.8, SD=.79) and post-test (M=4.47, SD=.76) scores; t(42)=-5.16, p=0.000, reveal improvement.

Student C had difficulty in the pre-test with basic grammar such as omitting verbs (“I very angry”), misuse of auxiliary verbs (“I listening it”), errors in subject-verb agreement, (“music help me”), and poor control over pronoun usage (“I want to they say”). In contrast, in the post-test, Student C demonstrated better control of verb tenses including simple present (“Hiroshima has a soccer team”), simple past (“I wanted to work part time”) and present perfect (“I have played soccer for 10 years”), while attempting simple future (“I’m going to find part time second semester”). In the pre-test, Student C predominately used simple sentences with a few instances of compound sentences, whereas in the post-test, the student attempted to use complex sentence structures, (“after I graduate this university, I want, I want to be teacher in primary school”). Basic grammatical problems still existed, but Student C had much better control over structures and syntax in the post-test.

4. Implications

Though shown to be statistically valid in showing both evidence of improvement of speaking skills and high inter-rater reliability (Stroupe & Panda, 2007), the 2004-2006 WLC Speaking Assessment, revisited in 2011, provided an opportunity to consider further improvements.

The scope of the 2004-2006 cycle of the WLC Speaking Assessment included more than 1000 students and was a strain on both the faculty and facilities. Both full- and part-time faculty members had been required to conduct many hours of assessments per day causing stress and fatigue. Some of the part-time faculty also felt that the responsibilities placed on them by the speaking assessments were beyond what
should be required. Because of the high number of assessments taking place simultaneously, the lack of appropriate assessment sites was sometimes an issue. These factors indicated that the scope of the assessment process needed revision.

Several recurring issues found in the format of the 2004-2006 WLC Speaking Assessment during the review process were noted. First, the amount of “interviewer-talk” was found to be excessive leading to less student production. Also, some interviewers tended to lead students through the interview or over-simplify questions without giving the student the chance to attempt to initiate speech. Furthermore, during a 5-min interview, it was difficult for some students to produce enough language to give an accurate sample of oral proficiency. In addition, students at higher proficiency levels were not pushed to the limit of their language abilities. One more formatting issue was the lack of consistency among the interviewers in the ways that they used the format protocols: for example, some interviewers deviated from the prescribed format. The persistence of these issues indicated the need for a change in the format of the speaking assessment.

In regards to the scoring rubric, some concerns with the previous version of the speaking assessment were noted. Some interviewers seemed to have criteria outside of the scoring rubric adopted for the assessment, i.e., willingness to talk, attitude, and previous interaction with the students, as part of the scoring system. Furthermore, comments during informal interviews with interviewers indicated some concerns with the scoring rubric. For example, some of the bandscale descriptors did not seem clear enough for the interviewers to differentiate between bands. Another concern was that the “Fluency and Coherence” component was too broad. Finally, pronunciation was not addressed in the scoring rubric.

Based on these concerns, a determination was made to revamp the WLC Speaking Assessment to mitigate some of the discrepancies between interviewers, and further refine the scoring rubric to eliminate any inconsistencies. The revised WLC Speaking Assessment is described below.

**Scope Changes**

When the WLC was asked to reinstate the WLC Speaking Assessment in the spring of 2012, it was decided that a limited sample cohort would be more appropriate than the previous population-based assessment to reduce demands made on faculty and facilities. With a sample cohort of roughly 200 students, each of the 18 full-time faculty would only be required to administer assessments for 4 to 5 hours during the assessment week and part-time faculty would not be required to participate. This would also alleviate the problem of locating appropriate venues.

**Assessment Format Changes**

**The pre-interview.**

During the pre-interview, the students watch a 2.5 to 4 min video clip during which they may take notes, and prepare to retell the story during the speaking assessment.

**The interview.**

**Part 1–Storytelling (2–3 min).**

The storytelling portion of the assessment gives the student the opportunity to produce language in a narrative form. The student should be given the freedom to speak uninterrupted until they finish retelling the story; however, the examiner may, if necessary, give
verbal encouragement to the student to continue speaking if the student speaks for less than 1 min and / or does not finish the story (i.e., Can you tell me anything else? What happened after...? What happened next?) or to ask for clarification (i.e., What do you mean by...? Could you tell me that again? Could you say that again?). The examiner should not volunteer new information during these prompts which the student has not produced previously. When the student has finished retelling the story, the examiner should ask the student to predict what might happen next (i.e., What do you think happens next in the story? What do you think would happen after that?).

This format change is intended to prevent too much “interviewer talk” while allowing for a longer, uninterrupted flow of narrative from the student. It also allows the student to feel some comfort and control of the speaking process because they are able to prepare and organize their ideas before the start of the assessment. This portion of the interview will allow the examiner time to preliminarily gauge the level of the students by paying particular attention to overall pronunciation, comprehension, comprehensibility, fluency, coherence, vocabulary and grammatical usage.

**Part 2: Topic discussion (5−7 min).**

The students will engage in a one-on-one interview with the examiner on a broad, familiar topic. The examiner offers the students two choices of general topics, i.e., sports, travel, hobbies, work, friends, etc., of which the students select one. This process will promote student ownership of the topic and encourage engagement, allowing for a more relaxed interchange during the interview. Throughout the exchange, the examiner should provide adequately challenging questions in which the student has the opportunity to utilize more complex discourse strategies, vocabulary and grammatical structures.

The two tasks incorporated into the revised WLC Speaking Assessment (Storytelling and Topic Discussion) follow a scaffolded format based on a spectrum of cognitive processes (Anderson & Krathwohl, 2001) as listed below:

1. **Remembering (Recalling information)** — Recognizing, listing, describing, retrieving, naming, finding
2. **Understanding (Explaining ideas or concepts)** — Interpreting, summarising, paraphrasing, classifying, explaining
3. **Applying (Using information in another familiar situation)** — Implementing, carrying out, using, executing
4. **Analysing (Breaking information into parts to explore understandings and relationships)** — Comparing, organising, deconstructing, interrogating, finding
5. **Evaluating (Justifying a decision or course of action)** — Checking, hypothesising, critiquing, experimenting, judging
6. **Creating (Generating new ideas, products, or ways of viewing things)** — Designing, constructing, planning, producing, inventing

The Storytelling section requires the use of remembering, understanding, and evaluating. The Topic Discussion section could potentially provide the interviewer the opportunity to utilize questions to evoke all six cognitive skills should the student be able to demonstrate sufficient oral proficiency for the complexity of the language demanded by the question. For example, the questions for the Topic Discussion on Technology illustrate this process.
1) Describe your favorite electronic item. What does it do? (Remembering and Understanding)

2) How has technology improved our lives? (Analyzing and Evaluating)

3) If you could design or redesign some piece of technology, what would it be? (Applying and Creating)

Similar to these examples, all prompts regardless of the topic are organized within this cognitive framework.

Scoring Rubric Changes

In the new format, the components for scoring were revised and the bandscale descriptors were clarified. A fourth component was added which includes task completion, comprehension, comprehensibility, and pronunciation. The task completion element of this component focuses on the ability of the student to effectively retell the story from the video clip and to what extent they are able to progress throughout the topic discussion. Also, this component considers the ability of the students to comprehend what the interviewer says. The interviewers will be asked to speak at a normal speed using the prepared scripted questions. The degree to which the students are able to comprehend language at normal speed or the vocabulary used would be indicators of the band in which the students fall. The comprehensibility of the students to the interviewer is also a part of this component. Finally, the pronunciation of the student, from phonemic to suprasegmental level, is evaluated in this component as well as the use of and reliance on L1 during oral production of English.

The revised Fluency and Coherence component now focuses on the ability of the students to express their knowledge of the topic in a coherent way including flow and ability to use language in appropriate situations. The use of discourse strategies, connected discourse, and the length of discourse are elements of this component as well. The focus of the Vocabulary and Grammar components is unchanged from the previous assessment scoring rubric; however, the descriptors for each of the bands have been clarified for easier use by the interviewer / raters.

Pilot of the New Format

In December 2011, the revised speaking assessment format was piloted with a small cohort of 32 students. Six faculty members participated in the pilot as interviewer / raters. Preliminary findings were positive, but a few minor changes in timing and logistics were identified. The most evident need recognized by the pilot was that more familiarity and training of the interviewer / raters with the scoring rubric was required for smooth and accurate assessments.

5. Conclusion

As this paper has demonstrated, the means by which to measure English oral proficiency remain unsettled (Bachman & Palmer, 1982; Lantolf and Frawley, 1985). Methodologies such as the format of the assessment, the speaking tasks administered, and the type of scoring (holistic or analytic) are also contested (Luoma, 2004).

Over the past ten years, the World Language Center has developed, tested and revised the speaking assessment instrument based on research developments in the field, feedback from interviewers, and observation of
countless hours of recorded interviews. As a result, significant changes to the format and the scoring rubric will be implemented. The aim of these changes is to increase uniformity in the assessment process and produce more reliable data.

Accountability in higher education institutions based on learning outcomes has become the global standard by which to measure the quality of educational institutions worldwide (Kushimoto, 2010). The process of developing and implementing the WLC Speaking Assessment informed the creation of a set of benchmarks (learning outcomes) utilized within WLC courses. Thus, the WLC Speaking Assessment has been an effective means by which to move the university towards greater accountability.

While this remains an ongoing process of reform and revision, it is our belief that the WLC Speaking Assessment is an effective way to measure students’ oral proficiency. In the increasingly competitive international higher education environment in which institutions are expected to demonstrate results based on learning outcomes, it is our hope that the WLC Speaking Assessment will help Soka University meet these expectations while simultaneously helping students improve English oral proficiency in order to increase their personal and professional opportunities.

References
Council for Higher Education Accreditation (CHEA). (2003). CHEA Institute for Research and Study of Accreditation and Quality As-
surance, 2003 Statement of mutual responsibilities for student 
learning outcomes: Accreditation, institutions, and programs. Re-


Language Assessment Quarterly, 3(1), 31-51.


Syndicate.
