The extent to which TOEFL iBT speaking scores are associated with performance on oral language tasks and oral ability components for Japanese university students

Gary Ockey, *Educational Testing Service*
Dennis Koyama, *Purdue University*
Eric Setoguchi, *University of California, Los Angeles*
Angela Sun, *Stanford University*

Available at: https://works.bepress.com/gary-ockey/7/
The Extent to which TOEFL iBT Speaking Scores are Associated with Performance on Academic Oral Language Tasks and Oral Ability Components of Japanese University Students

The Test of English as a Foreign Language internet-based Test (TOEFL iBT) includes a speaking section which aims to assess a test taker’s ability to communicate orally in an academic English environment (Xi, 2008). The test includes six items, which are all delivered by computer. Each item requires an asynchronous response; the expected response is based solely on an initial verbal and/or written stimuli. That is, after the prompt has been delivered by the computer, the test taker is expected to provide a response which will not receive a further reply. While these task types are reflective of some tasks that might be encountered in a real-world academic environment, such as responding to an instructor’s question, they are not reflective of others, such as a discussion among a group of peers, which requires synchronous communication. Given the importance of task authenticity in linking test scores and the target language use situation, the desired domain of score interpretation (Bachman and Palmer, 2010), it is plausible that TOEFL iBT speaking tasks do not effectively measure some components of academic oral ability, ones that are used during synchronous communication. It follows that research designed to determine the extent to which the TOEFL iBT asynchronous tasks validly assess all components of the ability to orally communicate in an academic environment is needed. To achieve this purpose, the present study investigated the extent to which scores on the TOEFL iBT speaking section were related to scores on oral tasks which have been used in EFL university contexts, including both synchronous and asynchronous tasks. Three tasks were selected for this purpose: a group oral discussion, a picture and graph description, and an oral presentation. These tasks were selected because they were designed to assess abilities needed for oral communication in English-medium university settings and used to assess L2 oral ability in

This is a manuscript of an article from Language Testing 32 (2015): 39, doi:10.1177/0265532214538014. Posted with permission.
programs preparing students for studying in these contexts. Scores on the TOEFL iBT speaking section were compared to scores on these three tasks and their component scores-- pronunciation, fluency, grammar/vocabulary, interactional competence, descriptive skill, presentation delivery skill, and answering questions--to determine which components of oral ability assessed by these tasks are best measured and which might be measured less well by the TOEFL iBT asynchronous speaking tasks.

**The oral ability construct**

According to Fulcher (2003), oral ability is generally thought to be comprised of knowledge and ability for use of phonology, which relates to pronunciation, stress, and intonation; accuracy of grammar, vocabulary, and cohesion; fluency, which refers to pauses, and repetition; strategic capacity, which should be considered as specific in nature to the context in which the language is used and includes the ability to manage communication; textual knowledge, which includes such abilities as turn taking and opening and closing gambits; and pragmatic and sociolinguistic competence specific to the speech act. This definition is in line with the speaking construct that TOEFL iBT aims to measure in which speaking was conceptualized as “the use of oral language to interact directly and immediately with others” in academic settings with the purpose of engaging in, acquiring, transmitting, and demonstrating knowledge (Jamieson, Eignor, Grabe, & Kunnan, 2009, p. 74).

**Development and research on TOEFL iBT speaking section scores**

The TOEFL iBT speaking tasks were developed based on a number of factors, including a needs analysis of the abilities required in a university environment. Rosenfeld, Leung, and Oltman (2001) investigated the extent to which certain abilities were perceived to be important in
a university context. Students (N = 345) and faculty (N = 370) from 21 North American universities judged various speaking abilities on a 1 to 5 scale based on their relevance to the needs of university students. Efforts were made to develop items that would assess the abilities judged to be most important by these academics.

A few studies have examined the extent to which the TOEFL iBT speaking section measures academic oral ability. Bridgeman, Powers, Stone, & Mollaun (2012) investigated the relationship between oral proficiency judgments made by untrained university undergraduate students and scores assigned for the speaking section of TOEFL iBT by ETS trained raters. Scores assigned by trained raters were compared to those assigned by 555 undergraduate students on 185 TOEFL iBT responses on two different test forms. Zero order correlations between the undergraduate judgments and the scores assigned by the TOEFL iBT trained raters were .79 for one test form and .67 for the other. The researchers concluded that the study provided strong evidence supporting the validity of TOEFL iBT scores as a measure of oral communicative competence.

Cumming, Grant, Mulcahy-Ernt, and Powers (2005) investigated the extent to which scores on the TOEFL iBT speaking section were representative of students’ performances in university ESL courses and assignments, including their in-class oral production. Seven “highly experienced” ESL teachers judged the authenticity of the tasks and the extent to which they elicited performances that corresponded to student performance in ESL classes. The researchers report that the majority of teachers felt that the tasks were well aligned with the domain of academic English required for studying in English-medium universities by marking 4 on a 1 to 5 scales with 5 indicating “Very Well” and 1 “Not Well”. Teachers also reported that student performance on the tasks were indicative of their performance in class activities. About 70% of
the performances were judged to be similar to classroom performance, 8% higher, and 22% lower.

Of particular relevance to the current research was a study conducted by Xi (2008), who investigated the extent to which TOEFL iBT speaking scores were related to locally developed teaching simulation tests designed to identify qualified international teaching assistants (ITA). The study included a comparison of TOEFL iBT speaking scores to the scores on four in-house university ITA assessments. Correlations between TOEFL iBT scores and component scores of the in-house tests varied widely. TOEFL iBT scores correlated most highly with grammatical competence measures (Bachman, 1990), such as presentation language and grammar/vocabulary, and least with task specific skills, such as teaching skill.

**University tasks**

A number of tasks have been used to assess academic oral ability in university contexts. Based on their own needs analysis and an understanding of oral assessment research and practice, the Japanese university selected for this study currently uses three tasks: a three-member group oral discussion, an individual picture and graph description, and a paired oral presentation.

**Group oral discussion task.** The group oral discussion task is increasingly used to assess oral ability, or more specifically, the ability to sustain a collaborative discussion, in second language classrooms. In this format, a small group of test takers are given a topic and then asked to have a discussion on the topic for a specified amount of time. The group oral discussion is becoming more commonly employed as a test task, in part, because of the use of small group discussions in English-medium universities, and consequently, in L2 classrooms (Van Moere, 2006). Kim (2006) found that participation in group discussions was frequently required in her
academic needs analysis. The group oral discussion task, along with communicative language teaching methodology, has become increasingly popular as a way of assessing the ability to discuss a topic in the classroom and in the real world, probably because of the importance of matching test tasks to real world tasks. A growing number of studies have provided evidence indicating that the group oral discussion task can be used to validly estimate a test taker’s oral ability (Bonk & Ockey, 2003; 2011; Liski & Puntanen, 1983; Van Moere, 2006).

**Picture and graph description task.** Visual prompts such as pictures and statistical graphs are often employed as a source of visual information for the test taker to comprehend and then describe (Xi, 2005). Description tasks make it possible to assess the ability to quickly comprehend, organize, and orally communicate information, which has been identified as important for functioning effectively in an academic setting (Ministry of Education, Culture, Sports, Science, and Technology-Japan: MEXT, 2009). In addition, the Common Core State Standards (CCSS: n.d.) adopted by many of the American states indicate that to be ready for college, students should be able to orally communicate about information which has been presented visually. Research, particularly on graph tasks, indicates that well specified tasks with careful attention to prompt design can lead to valid estimates of test takers’ oral abilities in an academic context (Xi, 2005, 2010).

**Oral presentations.** Oral presentations are commonly used to assess L2 ability in communicative classrooms, likely because they are reflective of tasks students perform in English-medium university contexts. Research indicates that university instructors and students consider formal presentations important parts of the academic curriculum. Ferris and Tagg (1996) surveyed 234 undergraduate and graduate course instructors and found formal presentations to be the third highest-ranking oral skill instructors expect students to master.
Ferris (1998) also surveyed 768 international students studying in US universities and discovered that formal speaking is perceived as the most important and the most problematic skill, and it was deemed by the respondents as the most worthwhile to develop in ESL classes. Orally presenting information is also indicated as important for college readiness in the CCSS (n.d.).

Research suggests that TOEFL iBT scores provide good indications of a test taker’s ability to orally communicate in an academic environment. Less is understood, however, about the extent to which the test measures components of academic oral ability for particular populations of test takers. It follows that research on particular TOEFL iBT test taker populations, such as Xi’s (2008) study on ITAs in the US, is needed to identify components of oral ability that are best and least well measured by TOEFL iBT asynchronous speaking tasks. Thus, the aim of this study was to examine the degree to which scores on the TOEFL iBT speaking section, which uses asynchronous tasks, assesses a broad construct of speaking ability for Japanese university students. The study addressed the following research questions.

**Research questions**

1. To what extent do scores on the TOEFL iBT speaking section relate to Japanese university students’ scores on an academic:
   1) impromptu L2 group oral discussion task?
   2) impromptu L2 picture and graph description task?
   3) prepared L2 oral presentation task?

2. To what extent do TOEFL iBT speaking section scores relate to Japanese university students’:
   1) pronunciation?
   2) fluency?
3) vocabulary/grammar?
4) interactional competence?
5) descriptive skill?
6) presentation delivery skill?
7) question answering skill?

Method

Participants

All participants (initial N = 226) who agreed to participate in the study were students enrolled in the English Department of a Japanese university specializing in foreign language study. A demographics survey indicated that the participants represented a somewhat diverse population of Japanese university students. The 226 students were from various regions of Japan. The majority were from Kanto, the most populous region of Japan, but in total 17 of the 46 Japanese prefectures were represented. Of the students in the study, 144 had studied in public high schools, and 62 in private high schools. Of the 226 students, 169 indicated that they hoped to study in a country where English is the primary language of use.

The participants were enrolled in a private university, which had complete autonomy over its curriculum. The university fostered a predominantly English-language instructional environment. Courses consisted of a communicative syllabus over the first two years followed by an additional two years of content courses. Like many of the university’s activities, the aim of these courses was preparing students for study in overseas English-medium universities. To this end, many of the required courses were conducted entirely in English, and formal Japanese-language instruction was limited to subject-specific coursework or seminars comprising less than 20% of courses. It should be note that the context of EFL study for this group of test takers
differs from that of most Japanese university students, who may spend only a few hours per week studying English formally and may have few opportunities to use English to orally communicate. Because an important aim of the study was to determine the extent to which TOEFL iBT asynchronous tasks validly assess all components of the ability to orally communicate in an academic environment, this group of test takers, who have formal and informal exposure to English, may be an ideal sample for the study.

Raters and rater training

University task raters. All university task raters were experienced EFL instructors who were teaching at the university. They were native speakers of English and held advanced degrees in applied linguistics or a related field of study. Eight raters were male and four were female. Six of the raters were from North America and six were from the United Kingdom.

To establish performance levels and provide an opportunity for rating practice for each of the three tasks, rater training was conducted according to established protocols, including demonstrations of task administration procedures by the researchers and viewings of recorded sessions from prior operational administrations of the tasks. Based on scores assigned by raters on a previous administration of the group oral discussion and the picture and graph description tasks, as well as the researchers’ judgments, a collection of recorded sessions representing a range of proficiencies of the population of students was selected and shown to raters as examples of levels of performance. After viewing each session, the rationale behind each proficiency rating was explained to the raters. The final stage of rater training consisted of ratings of videotaped sessions. A comparison of scores to the established benchmarks indicated that all raters were within one point of the established target scores. The training session took approximately two hours.
Rater training for the oral presentation task required more steps than for the other two university oral tasks because raters were not yet familiar with the use of the rating scale for this task. Prior to rater training for the oral presentation task, teachers selected a small group of willing students in their classes that were thought to represent the ability levels of the students in the study. The researchers then spoke with these students and their classroom teachers and identified six of the students that were judged to best represent the range of abilities in the target population. The six students were then assigned to pairs and given roughly 15 pages of information about an environmental issue—the 2010 Gulf oil spill. In a one-hour introductory session the six students were provided an explanation of a set of oral presentation task guidelines and time to ask any questions they had about the guidelines. An identical set of guidelines would eventually be given to all the participants in the study. At the end of the session, each pair was told that they would be paid for six additional hours of preparation and asked to return the following week to have their prepared oral presentations videotaped for the purpose of training raters. These recorded oral presentations were used in rater training sessions to demonstrate abilities associated with the different levels of the scoring rubrics.

**ETS TOEFL iBT speaking raters.** The TOEFL iBT speaking performances of the test takers were rated by ETS certified raters. The raters in the study all had extensive TOEFL iBT rating experience and were team leaders involved in training new ETS raters.

**Materials**

Five instruments were used in the study. Three were university oral tasks, namely, the group oral discussion, picture and graph description, and oral presentation, which were embedded in the university curriculum and designed to assess the academic oral abilities of the students in the study. The fourth measure was the TOEFL iBT speaking section, and the fifth
was a study log in which the students were required to record the amount of time they spent preparing for the oral presentation task.

Prior to the research reported here, all three university oral tasks had been used to measure the English oral abilities of students at the university. The group oral and the picture and graph tasks had established scoring rubrics prior to the study while a unified scoring procedure had not been adopted for oral presentations. The three tasks were developed based on a needs analysis of Japanese students studying in an academic English environment and were designed to align with the students’ needs and the program’s objectives and curriculum. The purpose of these oral tasks is to provide an objective measure of academic oral proficiency. Scores are used to aid in assigning course grades (20% of the overall grade) and to place students into proficiency tracks in subsequent years of study. Thus, while the tasks are designed to align with the needs of the students and the curriculum, they are also meant to be measures of L2 oral proficiency.

Each university oral task is designed to assess various components of oral language ability: pronunciation, fluency, vocabulary/grammar, and interactional competence, descriptive skill, or presentation delivery skill and question answering skill. To achieve this purpose, analytical rating scales which share identical scales for pronunciation, fluency, and vocabulary/grammar are used to judge oral language ability as demonstrated by performance on each task. Each rating scale also includes a task-specific scale, and in the case of the oral presentation task, a fifth scale devoted to a short, impromptu question and answer session. For each subscale, a nine-point scoring system is used with a minimum score of 1.0 and a maximum of 5.0. Whole number scores within a proficiency band indicate that test takers partially demonstrate the ability as described, whereas half-point scores indicate that test takers have completely demonstrated the ability described in a given category. On each of the university
tasks, raters assign scores on all oral ability components in real time during the students’ performances. Students are assigned the average score assigned by the two raters. The scales were developed by an institutionally-based assessment team and had been revised and adapted from time to time prior to this study. Appendix A shows the three scoring scales: pronunciation, fluency, and vocabulary/grammar, which are common to the three university tasks. Appendix B provides each of the three task-specific scales.

**University group oral discussion task.** The group oral discussion task is designed to assess test takers’ abilities to sustain a discussion on an assigned academically-themed topic. In the task, three students who are studying in the same classes are assigned to discuss a topic for eight minutes. Each group is randomly assigned one of five prompts. An example prompt is:

Agree or disagree with the following statement: The most important thing to do for job hunting is to study abroad for one year. Use specific examples to support your position. Try to convince the group that your position is best.

Students are invited into a testing room and asked to sit in desks that have been arranged in a small circle. They are each given a prompt written in Japanese and English, which is also read aloud by one of two raters. After 30 seconds, one of the raters collects the prompts and says, “Please begin the discussion.” Two trained raters sit outside of the group and assign scores for pronunciation, fluency, vocabulary/grammar, and interactional competence. Interactional competence has been defined in various ways (Fulcher, 2003; McNamara, 1997) but for the purpose of this study, it relates to participation and smoothness of interaction (e.g., turn-taking, responding to others, asking questions, introducing new gambits, paraphrasing, and hedging). After eight minutes, one of the raters ends the test by saying, “That’s eight minutes. Thank you.”

**University picture and graph description task.** The picture and graph description task includes two subtasks, describing a picture and describing a statistical graph, both of which are
designed to assess the ability to process and describe visual prompts in an impromptu setting. Students are invited into a small testing room and asked to sit at a table. Two raters administer the test and sit on the side of the table opposite from the test taker. For the first task, the test taker is given 60 seconds to describe a picture. The picture prompt is then removed from the desk and the test taker is given a graph and 60 seconds to describe it. An example of each task is provided in Appendix C. Based on the two-minute performance, the two trained raters assign scores for pronunciation, fluency, vocabulary/grammar, and descriptive skill. The descriptive skill band relates to the ability to orally describe a visual input with organization, thoroughness, and precision (See Appendix B). Raters are trained to give equal weighting to the two tasks when assigning a score. One of six test forms is randomly used to ensure that test takers who take the test after others are not advantaged by having information about the prompts.

**University academic oral presentation task.** The oral presentation task is designed to assess the ability to prepare and deliver an organized presentation and to respond to spontaneous questions about its content. Students present in pairs, and each pair is given eight minutes to complete the task. To prepare the students, a video of a model presentation is shown and discussed in class at the beginning of the school semester. The model presentation is given by two senior students who had been consistently placed into the highest level of proficiency on the university’s oral placement examination. Three times each semester, students at the university give presentations on academic topics. The second presentation given in the semester was used as the presentation task in the study. Therefore, all students had completed at least one presentation and were familiar with the procedures at the time of the presentation task used in the study.
The presentation task topic is introduced approximately three weeks prior to the date of the presentation task and is loosely related to the content that students have been studying in their English classes. Students are randomly assigned to pairs and in this case were asked to present on one of two topical areas, the media or the environment. Presentation task guidelines require that each pair use one piece of poster paper, which is provided by the teacher to ensure uniformity in color and size. The poster paper is expected to display the title of the presentation and one or more pictures to help make the presentation more transparent. Guidelines mandate that students cannot use scripts when giving their presentations. All students are given two 90-minute classes to prepare, but teachers provide no guidance or feedback during these sessions. The students are encouraged to prepare additionally outside of class and keep a log of their preparation time.

Two trained raters, the presenters’ teacher, and roughly 10 classmates are in attendance when test takers complete the presentation task. After the prepared portion of the presentation task, raters are given time to complete their ratings for pronunciation, fluency, vocabulary/grammar, and presentation delivery skill. The presentation delivery skill band relates to content development, logic of organization, and effectiveness and appropriateness of delivery, including eye contact, gestures, use of visual aids, audience awareness, and register (See Appendix B). Finally, the teacher asks each presenter one non-rehearsed question about the content of the presentation, and the raters provides a score for the question and answer band of the score scale, which aims to assess a test taker’s comprehension of the question and ability to provide an organized response which addresses the question (see Appendix B). Raters are instructed to refrain from taking into consideration performance on the prepared portion of the
presentation when assigning scores for the question and answer band. They are also asked to not change ratings previously given to the prepared presentation portion of the task after listening to the question and answer session. Although the presentations are given in pairs, each student is assigned an individual score for each of the five rating scales.

TOEFL iBT speaking section. The computer-delivered speaking section of the TOEFL iBT has six tasks designed to measure the ability to orally communicate in an academic context. The first two tasks, which are designed to assess only speaking, require test takers to give their opinion about everyday topics, while the latter four are integrated tasks which aim to measure speaking integrated with listening and/or reading. Two of these latter tasks are about campus situations and two are about academic content. For these tasks, test takers are expected to summarize and/or provide their opinion on an academic topic which is introduced through a short reading and/or listening stimulus. The ETS certified raters who participated in the study are trained to consider three components of oral ability when assigning scores: delivery, which includes the pace and clarity of speech; language use, which relates to the effectiveness of use of grammar and vocabulary; and topic development, which is the depth and coherence of ideas. Performance on each of the six tasks is assigned an independent score on a five-point scale. The TOEFL iBT speaking rating scale is available at: http://www.ets.org/Media/Tests/TOEFL/pdf/Speaking_Rubrics.pdf. For the study, students were randomly assigned one of two research forms of the test. The research forms are retired forms of the operational test that are not publicly available.

Each performance was judged by two raters, and no rater scored more than two of the six items on the test. The two scores assigned for each of the six tasks were averaged, except in
cases in which the scores were more than one point apart. These cases were adjudicated by a third rater, and the score assigned by the third rater was used as the final score. In all cases, this third score was either equivalent to one of the other assigned scores or between the two assigned scores.

**Presentation task study logs.** Students were required to keep a record of the time they spent preparing for the oral presentation task outside of class. These logs were meant to help students reflect on the activities and the time that they spent preparing and to provide the researchers with a control for study time in the analyses.

**Procedure**

All students participating in the study completed the three university oral tasks as a required part of the curriculum and a research version of the TOEFL iBT speaking section under conditions similar to those required for operational TOEFL iBT administrations. Students were informed that 20% of the term grades would be based on their performances on the three tasks and the TOEFL iBT speaking section. To ensure that the study was ethical, all participants were given the choice of opting out of taking the TOEFL iBT speaking section. No students took this option, however, possibly because they were interested to see how they would perform on the test and/or because they were compensated for participating in the study.

The university oral tasks were administered between 10 and 14 weeks into the 15-week semester. Prior to completing the group oral discussion and the picture and graph description tasks, participants practiced by performing simulated tasks and receiving feedback from the teacher and classmates during regular classes. Approximately two weeks before the test was administered, a 90-minute introduction to the TOEFL iBT speaking section was provided in regular classes. The session included an introduction to the test and practice and feedback on
each of the six tasks. The ETS practice materials available online at: http://www.ets.org/toefl/ibt/prepare/sample_questions/ were used during the practice session.

Each student completed the three university tasks and the TOEFL iBT speaking section within a two-week period, and it took approximately four weeks to complete data collection for all students. To avoid an effect for order of task administration, a counter-balanced design following a Latin square approach was employed (Keppel & Wickens, 2004). Due to technical difficulties, the recordings of 30 test takers’ TOEFL iBT sound files were determined to not be ratable. All 30 of these test takers were asked to take the TOEFL iBT research form that they had not previously taken. The retests took place approximately three weeks after the first test administration, and students were not informed that they would need to take the retest until almost immediately before the test was administered. Given that all students had taken practice tests and prepared for the TOEFL iBT speaking section prior to the test, it was determined that a test effect was unlikely for these 30 test takers. After retests were included, the scores of only four students were excluded due to technical issues in the scoring of the TOEFL iBT speaking section. No student had to be retested for any of the university oral tasks, and complete scores were available for all. The scores of 222 of the 226 original participants were used in all analyses in the study.

Results

Descriptive statistics and estimates of reliability

Descriptive statistics and estimates of reliability are provided for the group oral discussion task in Table 1, the picture and graph description task in Table 2, and the oral presentation task in Table 3. Score reliability of the university oral tasks was estimated, treating each rating as a score in the Cronbach Alpha formula (Carr, 2011).
Table 1 Descriptive Statistics and Reliability Estimates of the Group Oral Discussion Task

<table>
<thead>
<tr>
<th></th>
<th>Pronunciation</th>
<th>Fluency</th>
<th>Lexis/</th>
<th>Interactional</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Mean</strong></td>
<td>3.38</td>
<td>3.41</td>
<td>3.35</td>
<td>3.75</td>
<td>13.89</td>
</tr>
<tr>
<td><strong>Max</strong></td>
<td>5.00</td>
<td>4.75</td>
<td>5.00</td>
<td>5.00</td>
<td>18.50</td>
</tr>
<tr>
<td><strong>Min</strong></td>
<td>1.75</td>
<td>2.00</td>
<td>2.00</td>
<td>1.25</td>
<td>9.25</td>
</tr>
<tr>
<td><strong>SD</strong></td>
<td>0.57</td>
<td>0.62</td>
<td>0.50</td>
<td>0.59</td>
<td>1.99</td>
</tr>
<tr>
<td><strong>Skewness</strong></td>
<td>0.23</td>
<td>0.05</td>
<td>0.16</td>
<td>-0.61</td>
<td>-0.05</td>
</tr>
<tr>
<td><strong>Kurtosis</strong></td>
<td>0.00</td>
<td>-0.62</td>
<td>0.24</td>
<td>0.87</td>
<td>-0.40</td>
</tr>
<tr>
<td><strong>Alpha</strong></td>
<td>.65</td>
<td>.74</td>
<td>.71</td>
<td>.71</td>
<td>.78</td>
</tr>
</tbody>
</table>

Table 2 Descriptive Statistics and Reliability Estimates for Picture and Graph Description Task

<table>
<thead>
<tr>
<th></th>
<th>Pronunciation</th>
<th>Fluency</th>
<th>Vocabulary/</th>
<th>Descriptive</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Mean</strong></td>
<td>3.35</td>
<td>3.34</td>
<td>3.29</td>
<td>3.38</td>
<td>13.37</td>
</tr>
<tr>
<td><strong>Max</strong></td>
<td>5.00</td>
<td>5.00</td>
<td>4.50</td>
<td>4.50</td>
<td>18.25</td>
</tr>
<tr>
<td><strong>Min</strong></td>
<td>2.25</td>
<td>2.00</td>
<td>2.00</td>
<td>2.25</td>
<td>9.00</td>
</tr>
<tr>
<td><strong>SD</strong></td>
<td>0.51</td>
<td>0.58</td>
<td>0.45</td>
<td>0.45</td>
<td>1.71</td>
</tr>
<tr>
<td><strong>Skewness</strong></td>
<td>0.24</td>
<td>0.05</td>
<td>0.14</td>
<td>-0.10</td>
<td>0.09</td>
</tr>
<tr>
<td><strong>Kurtosis</strong></td>
<td>0.06</td>
<td>-0.09</td>
<td>0.01</td>
<td>-0.11</td>
<td>0.02</td>
</tr>
<tr>
<td><strong>Alpha</strong></td>
<td>.62</td>
<td>.79</td>
<td>.64</td>
<td>.64</td>
<td>.81</td>
</tr>
</tbody>
</table>

Table 3 Descriptive Statistics and Reliability Estimates for Oral Presentation task

<table>
<thead>
<tr>
<th></th>
<th>Pronunciation</th>
<th>Fluency</th>
<th>Vocabulary/</th>
<th>Presentation</th>
<th>Q and A</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Mean</strong></td>
<td>3.39</td>
<td>3.62</td>
<td>3.64</td>
<td>3.59</td>
<td>3.30</td>
<td>17.52</td>
</tr>
<tr>
<td><strong>Max</strong></td>
<td>5.00</td>
<td>5.00</td>
<td>5.00</td>
<td>4.75</td>
<td>5.00</td>
<td>24.50</td>
</tr>
<tr>
<td><strong>Min</strong></td>
<td>1.50</td>
<td>1.75</td>
<td>2.25</td>
<td>1.25</td>
<td>1.50</td>
<td>9.75</td>
</tr>
<tr>
<td><strong>SD</strong></td>
<td>0.65</td>
<td>0.71</td>
<td>0.52</td>
<td>0.63</td>
<td>0.73</td>
<td>2.62</td>
</tr>
<tr>
<td><strong>Skewness</strong></td>
<td>-0.13</td>
<td>-0.36</td>
<td>-0.01</td>
<td>-0.73</td>
<td>0.08</td>
<td>-0.18</td>
</tr>
<tr>
<td><strong>Kurtosis</strong></td>
<td>0.23</td>
<td>-0.12</td>
<td>-0.22</td>
<td>0.95</td>
<td>-0.49</td>
<td>0.13</td>
</tr>
<tr>
<td><strong>Alpha</strong></td>
<td>.64</td>
<td>.73</td>
<td>.56</td>
<td>.64</td>
<td>.69</td>
<td>.72</td>
</tr>
</tbody>
</table>
As can be seen in Tables 1, 2, and 3, means were near 3.5 on all of the university tasks, and standard deviations, skewness, and kurtosis values suggested normal distributions for all. Reliability estimates for the group oral discussion and the picture and graph description tasks were slightly higher than for the oral presentation task.

The descriptive statistics for the presentation task study logs are presented in Table 4.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th>Min</th>
<th></th>
<th></th>
<th>Max</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>Max</td>
<td>Min</td>
<td>SD</td>
<td>Skewness</td>
<td>Kurtosis</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Presentation task study logs</td>
<td>12.18</td>
<td>43.00</td>
<td>1.67</td>
<td>6.61</td>
<td>1.53</td>
<td>3.97</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The amount of time that students spent studying outside of class varied widely. A standard deviation of 6.61 confirms this large difference in time spent preparing for the oral presentation task.

The descriptive statistics for the TOEFL iBT speaking section are presented in Table 5.

<table>
<thead>
<tr>
<th>TOEFL iBT Speaking</th>
<th>Item 1</th>
<th>Item 2</th>
<th>Item 3</th>
<th>Item 4</th>
<th>Item 5</th>
<th>Item 6</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>2.00</td>
<td>2.18</td>
<td>1.77</td>
<td>1.63</td>
<td>1.84</td>
<td>1.63</td>
<td>11.05</td>
</tr>
<tr>
<td>Max</td>
<td>3.50</td>
<td>3.50</td>
<td>3.00</td>
<td>3.00</td>
<td>3.00</td>
<td>3.00</td>
<td>17.00</td>
</tr>
<tr>
<td>Min</td>
<td>0.00</td>
<td>1.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>4.00</td>
</tr>
</tbody>
</table>
The scores are fairly well centered on the 0 to 4 scale, with means ranging from 1.63 to 2.18. No students achieved the maximum score on any of the six items, suggesting that the test takers did not include any high-advanced individuals. Reliability as estimated by Cronbach’s Alpha was .82. Estimates were based on treating each item as a score.

**Relationship between oral presentation task scores and preparation time**

Given that participants were encouraged to spend extra time outside of class preparing for the oral presentation task, variation in study time was taken into account when investigating the relationship between TOEFL iBT speaking scores and scores on the oral presentation task. Correlations between oral presentation task preparation time and scores on the oral presentation task are displayed in Table 6.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pronunciation</td>
<td>.04</td>
</tr>
<tr>
<td>Fluency</td>
<td>.06</td>
</tr>
<tr>
<td>lexis/grammar</td>
<td>.11</td>
</tr>
<tr>
<td>Presentation delivery skill</td>
<td>.08</td>
</tr>
<tr>
<td>Question &amp; Answer</td>
<td>.06</td>
</tr>
<tr>
<td>Total</td>
<td>.08</td>
</tr>
</tbody>
</table>

* indicates significant at the .05 level

As can be seen, all correlations between oral presentation task preparation time and scores on the oral presentation task were small and positive. None, however, were significant. Given that no
significant relationship was found between preparation time for the oral presentation task and oral presentation task scores, preparation time was not modeled in the analyses.

**Correlations among scores on the TOEFL iBT Speaking test and university oral tasks**

Correlational analyses corrected for attenuation (Carr, 2011), due to imperfect reliability of the assessments, was used to provide an indication of the degree to which TOEFL iBT scores were related to scores on the university oral tasks. When assessments have measurement error, as was the case in this study (reliabilities ranged between .62 and .82), the measurement error leads to correlations between the variables measured by these assessments that are lower than the true correlations among the abilities of interest. Thus, correcting correlations for attenuation likely provides a better estimate of the correlations among the abilities under study\(^1\). It was therefore determined that such corrections were appropriate for this study. The corrected for attenuation, that is, disattenuated correlations among the TOEFL iBT speaking section scores and the scores on the university oral tasks are presented in Table 7. Correlations between the TOEFL iBT speaking scores and the university oral task scores range from a high of .76 for the group oral discussion to a low of .68 for the oral presentation. Scores on the picture and graph description task correlated with scores on the TOEFL iBT speaking section at .73. The correlations among the university oral tasks were between .67 and .73.

| Table 7 Correlations between TOEFL iBT Speaking and university oral tasks |
|-----------------------------|-----------------|-----------------|
| TOEFL iBT Speaking          | Group Oral Discussion | Picture & Graph Description |

\(^1\) When measurement error is not random, disattenuation can lead to inflated correlations and should always be considered when the technique is used.
Correlations have been corrected for attenuation
* indicates significant at the .05 level.

### Correlations among scores on the TOEFL iBT Speaking test and components of university oral tasks

Disattenuated correlations between TOEFL iBT speaking scores and each of the components of oral ability are presented in Table 8. TOEFL iBT scores showed moderate to high correlation with all oral ability component scores of the university tasks.

<table>
<thead>
<tr>
<th></th>
<th>Group Oral Discussion</th>
<th>Picture &amp; Graph Description</th>
<th>Oral Presentation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pronunciation</td>
<td>.71*</td>
<td>.70*</td>
<td>.63*</td>
</tr>
<tr>
<td>Fluency</td>
<td>.74*</td>
<td>.68*</td>
<td>.59*</td>
</tr>
<tr>
<td>Lexis/grammar</td>
<td>.73*</td>
<td>.75*</td>
<td>.50*</td>
</tr>
<tr>
<td>Interactional competence</td>
<td>.63*</td>
<td></td>
<td>.61*</td>
</tr>
</tbody>
</table>

**Table 8 Correlations between TOEFL iBT speaking scores and components of university oral task scores**

Correlations have been corrected for attenuation
* indicates significant at .05 level
Correlations above .50 are generally associated with large effect sizes (Rubin, 2012). These correlations, corrected for attenuation, ranged from a high of .74 to a low of .51. A similar pattern of relationships exists between TOEFL iBT speaking scores and the four components of oral ability measured by the group oral discussion and the picture and graph description tasks. Scores on interactional competence of the group oral discussion and the descriptive skill of the picture and graph description tasks were substantially less related to TOEFL iBT scores than pronunciation, fluency, and vocabulary/grammar. This finding was similar to that of oral ability component scores yielded from the oral presentation task. The only difference was that the vocabulary/grammar oral ability component score was lower than the fluency and pronunciation component scores. Correlations between TOEFL iBT speaking scores and oral ability component scores of the oral presentation task were generally lower than the correlations between TOEFL iBT speaking scores and the oral ability component scores of the group oral discussion and the picture and graph description tasks. The question answering oral ability component score yielded from the oral presentation task was associated to roughly the same degree as the pronunciation and fluency component scores.

Discussion

To determine the degree to which scores on the TOEFL iBT speaking section, which uses asynchronous tasks, assesses a broad construct of speaking ability for Japanese university students, their relationship with three university oral tasks and their components were investigated. First, however, it should be noted that no significant relationship was found between the oral presentation task scores and the amount of time spent preparing for the presentation task, suggesting that the university tasks, at least the oral presentation task, which was arguably the most likely to be affected by preparation, largely measured academic oral
proficiency as was intended in the study. The first research question in the study aimed to determine the extent to which scores on the TOEFL iBT speaking section were associated with performance on the group oral discussion, picture and graph description, and oral presentation tasks.

Based on correlational analyses corrected for attenuation, it was found that TOEFL iBT speaking scores were highly related to performances on all three tasks: group oral discussion, \( r = .76 \), picture and graph description, \( r = .73 \), and oral presentation task, \( r = .68 \), suggesting that the TOEFL iBT does measure a broad construct of speaking ability. The results, however, point at components of oral ability that the asynchronous TOEFL iBT speaking tasks measure more effectively than others. Given that the oral presentation task was the only prepared task, the findings suggest that scores on the TOEFL iBT speaking section may be more associated with unprepared than prepared tasks, although the difference in the size of these correlations was minimal, suggesting that such an effect is quite small. This finding likely stems from the fact that little preparation time is given before test takers are expected to complete the TOEFL iBT tasks. This may provide some support for the aim of TOEFL iBT to assess the ability to orally communicate “immediately” (Jamieson, Eignor, Grabe, & Kunnan, 2009, p. 74).

The differences in the types of tasks that the students completed for the TOEFL iBT speaking section and the university oral tasks likely account for some of the unexplained variance in oral task scores. Four of the six TOEFL iBT speaking tasks differ from the university oral tasks in that they are integrated with reading and listening. Test takers listen to and/or read a stimulus to which they are expected to orally respond. The group oral discussion task requires test takers to listen to each other, but such synchronous communication, or co-construction of meaning in real time, provides them with opportunities to ask for repetition and clarification if
they do not understand. Thus, the asynchronous TOEFL iBT tasks likely measure reading and listening as well as speaking, while the university oral tasks focus more on speaking (and listening for the group oral discussion).

The research also aimed to determine the extent to which TOEFL iBT speaking scores were related to pronunciation, fluency, vocabulary/grammar, interactional competence, descriptive skill, and presentation delivery skill. To shed light on this issue, a correlational analysis corrected for attenuation was conducted between TOEFL iBT speaking scores and scores on each of these components of oral ability. The results of the analysis were consistent between the group oral discussion and the picture and graph description tasks, but slightly different for the oral presentation task. For the group oral discussion and the picture and graph tasks, scores for pronunciation, fluency and vocabulary/grammar were similarly related, while interactional competence and descriptive skill were somewhat less associated with TOEFL iBT speaking scores than pronunciation, fluency, and vocabulary/grammar. The results were similar for the oral presentation task, except the vocabulary/grammar oral ability component was least related to TOEFL iBT speaking test scores. The question answering oral ability component was related to TOEFL iBT speaking scores similarly to the pronunciation and fluency components. The findings suggest that TOEFL iBT speaking scores measure interactional competence, descriptive skill, and presentation delivery skill to a lesser extent than they do the oral ability components of pronunciation, fluency, and vocabulary/grammar common to the three oral ability tasks. These findings are in line with Xi’s (2008), which indicated that the TOEFL iBT measured grammatical competence to a larger extent than task specific skills, such as teaching ability. This finding raises the question of the extent to which interactional competence, descriptive skill, and/or presentation delivery skill are components of the ability to orally communicate in an
academic environment and by extension the abilities that the TOEFL iBT speaking section should assess, as elaborated on below.

Fulcher’s (2003) definition of the oral ability construct includes knowledge of turn-taking and opening and closing gambits, which are important aspects of interactional competence, and Butler, Eignor, Jones, McNamara, & Suomi (2000) argued that the ability to interact directly and immediately with others was important when they defined the oral ability construct that TOEFL should measure. Thus, the interactional competence component of oral ability is clearly a crucial part of the TOEFL speaking construct.

It may not be surprising that TOEFL iBT speaking scores were related to the interactional competence component of oral ability to a substantially smaller extent than to pronunciation, fluency, and vocabulary/grammar. While the iBT speaking score may be a good indicator of a test taker’s ability to listen to, process, and summarize orally the information presented by others, it would not be expected to capture the test taker’s ability to negotiate meaning with others by asking clarification or follow-up questions. This may be, at least in part, due to the asynchronous nature of the TOEFL iBT speaking tasks: the expected responses are based only on an initial verbal and/or written stimulus. Thus, it can be argued that the TOEFL iBT asynchronous speaking tasks may not effectively assess these abilities to orally communicate in real time, at least not to the same extent that they measure other components of speaking ability. This finding may also be attributable to the TOEFL iBT speaking rating scales as compared to the university task rating scales. The TOEFL iBT delivery and language use scales are similar to the university tasks’ pronunciation, fluency, and vocabulary/grammar scales, but there is no scale that is designed to assess interactional competence on the TOEFL iBT. Hence, the study indicates that in addition to possible task revision, it may be helpful to revise the rating scales to
include aspects of interactional competence. A practical implication of the findings is that test users, when interpreting iBT speaking scores and making judgments about test takers’ academic oral English abilities, take into consideration the interactional aspects of the speaking construct the scores may or may not reflect.

The descriptive skill oral ability component of the picture and graph description task aims to measure the “effectiveness of oral communication of key information in a visual prompt (organization, thoroughness, and precision).” The ability to identify important information and describe objects has been identified as important for communicating in the Japanese classroom (MEXT, 2009), and the ability to orally discuss visual information has been recognized as important for college readiness in the United States. It may therefore be argued that this ability be assessed by the TOELF iBT speaking section. On the other hand, it could be argued that this more task specific oral ability may not be as crucial to the ability to orally communicate in an academic environment as the more general abilities of pronunciation, fluency, and vocabulary/grammar which would be essential for performing any oral task. A weaker relationship between TOEFL iBT speaking scores and description skills than between TOEFL iBT scores and these more general abilities may be appropriate.

The presentation delivery skill component of the oral presentation task is designed to assess the degree to which the “presentation is developed in content, logically organized, and academically delivered (eye contact, gestures, use of visual aids, audience awareness, register).” It may not be surprising that these skills are not measured very well by a test such as the TOEFL iBT’s speaking section for which only audio files are used for rating performance, since it is not possible to assess abilities such as eye contact and gestures. It could be argued that this non-verbal information should be included in the academic oral ability construct. The impact of
nonverbal features on communication is reflected in the results of other research on assessing speaking. Seddon and Pedrosa (1990) observed that when scoring oral science tests, examiners who interacted with a student face-to-face tended to give higher scores compared to examiners who only listened to audio of the same student’s response. Jenkins and Parra (2003) examined the effect of examinees’ nonverbal behavior on raters’ perception of their language abilities and concluded that linguistically weaker examinees who employed nonverbal strategies to negotiate meaning with evaluators received higher scores than other borderline examinees. The findings of these studies indicate that nonverbal features play powerful compensatory roles in performance on oral test tasks and underscore the importance of considering their inclusion in an oral ability construct.

It is important to note that conducting the study at a single institution in Japan likely limits the generalizability of the findings. The TOEFL iBT speaking section scaled scores for the participants ranged from 22 to 5, with an average of 14 and a standard deviation of 3.3\(^2\). Both the mean and standard deviation are lower than the Japanese general test taker population which was 17 and 4.6, respectively, in 2012 (Educational Testing Service, 2013). It might therefore be expected that because of this somewhat restricted range, correlations between TOEFL iBT speaking scores and the university oral tasks could be depressed in the study.

Overall, the findings of the study provide support for the claim that the TOEFL iBT asynchronous speaking tasks assess the ability to communicate orally in an academic environment. The research also suggests the need for institutions interpreting TOEFL iBT speaking scores to consider the extent to which the scores reflect different aspects of academic

\(^2\)The scores shown in Table 5 were scaled to the 30-point TOEFL iBT reporting scale to make them comparable to TOEFL iBT reported scores. ETS provides a brief description of this scaling at: http://www.ets.org/s/toefl/pdf/94227_unlweb.pdf
oral English skills. Moreover, to better assess the academic oral ability construct that TOEFL iBT aims to assess, the research suggests the need to consider the addition of tasks to the TOEFL iBT speaking section that more adequately assess interactional competence. It should be noted, however, that current techniques used to assess interactional competence, such as real-time paired and group discussions have been shown to have their own limitations (e.g., Galaczi, 2008; Ockey, 2009; Van Moere, 2006) that might not make them appropriate for high stakes assessments such as TOEFL iBT. The research also raises questions about whether TOEFL iBT tasks should be redesigned to assess the ability to orally respond to visual stimuli, and whether or not video files rather than only audio files should be used when rating speaking.
References


## Appendix A Pronunciation, Fluency, and Vocabulary/Grammar scoring scales

<table>
<thead>
<tr>
<th>Pronunciation</th>
<th>Fluency</th>
<th>Vocabulary / Grammar</th>
</tr>
</thead>
<tbody>
<tr>
<td>● Word level</td>
<td>● Automatization: ability to formulate utterances quickly and speak smoothly</td>
<td>● Correct grammatical form</td>
</tr>
<tr>
<td>● Sentence Level: ability to ‘blend’ or link sound within or between words</td>
<td>● Speaking speed</td>
<td>● Suitability of vocabulary</td>
</tr>
<tr>
<td>● Stress, rhythm, and intonation</td>
<td>● Hesitations and pausing</td>
<td>● Displaying ability to use (or attempting to use) different grammatical structures and vocabulary suitably in context</td>
</tr>
<tr>
<td>● Accent</td>
<td></td>
<td>● Collocations and correct word choice</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>1 ~ 1.5</th>
<th>2.0 ~ 2.5</th>
<th>3.0 ~ 3.5</th>
</tr>
</thead>
</table>

### Pronunciation

- **Very heavy accent, that would lead to a breakdown in communication**
- **Only uses accented phonology and rhythm; words not blended together**
- **Fragments of speech**
- **Halting, often incomprehensible**
- **Communication nearly impossible**
- **No evidence of grammar knowledge**
- **Knows few words, and uses them in isolation**
- **Unable to share simple ideas**
- **Communication not possible**

### Fluency

- **Slow strained, unnatural speech**
- **Frequent unnatural groping for words**
- **Long unnatural pauses**
- **Communication difficult**
- **Some very limited grammar knowledge evident**
- **Limited vocabulary but inexpert usage**
- **Little or no attempt at complex vocabulary or grammar**
- **Ideas can be shared, but with likely comprehension difficulties**

### Vocabulary / Grammar

- **Has not mastered some difficult sounds of English, but should be mostly understandable to interlocutors**
- **Makes regular attempts to blend words but may still stress words incorrectly**
- **Speech is hesitant; somewhat unnatural**
- **Unnatural groping for words and unfilled spaces may persist, but it does not impede communication**
- **May overuse fillers, or demonstrate other unnatural usages**
- **Overly reliant on a small range of simple grammar and vocabulary to express ideas**
- **Shows little or no evidence of ability to control difficult grammar or vocabulary**

<table>
<thead>
<tr>
<th>4.0 ~ 4.5</th>
<th>5.0</th>
</tr>
</thead>
</table>

### Pronunciation

- **May not have mastered all the sounds of English, but has good control of sentence stress and intonation**
- **Accent does not interfere with comprehension; can blend words consistently**
- **Occasional misuse of fillers and lexical chunks.**
- **Groping and frequent repair may still be evident, but is not overly distracting to listeners**
- **Shows evidence of ability to control difficult grammar or vocabulary and attempts to use a range of forms**
- **May continue to make mistakes, but should be comprehensible**

### Fluency

- **Conversation should proceed smoothly, with little impediment**
- **Uses fillers, markers, lexical chunks effectively**
- **Groping may occur, but seems natural & fluent**
- **Demonstrates excellent control of a range of grammar and vocabulary**
- **Mistakes may still occur, but these should not impede meaning**
- **Chunked lexical items, such as idioms and collocations may be present and used correctly**
<table>
<thead>
<tr>
<th>Interactional competence (Group oral discussion)</th>
<th>Descriptive Skill (Picture and graph description)</th>
<th>Presentation delivery skill (Oral presentation)</th>
<th>Question answering (Oral presentation)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Participation and smoothness of interaction (turn-taking, responding to others, asking questions and introducing new gambits, paraphrasing, hedging)</td>
<td>Effective communication of key information in a visual prompt (organization, thoroughness, precision)</td>
<td>Presentation is developed in content, logically organized, and academically delivered (eye contact, gestures, use of visual aids, audience awareness, register)</td>
<td>Clarifying question, giving general and specific answers, checking response, providing clear detailed answer</td>
</tr>
</tbody>
</table>
| **1.0 ~ 1.5** | *Does not initiate interaction*  
*Uses mostly single word and simple short phrases*  
*Minimal or nonexistent responses to others* | *Shows no organization of information*  
*Unable to communicate any significant information about the visual prompt*  
*Description contains little or no accurate information about the visual prompt* | *May not understand question.*  
*May respond, but does not address question.*  
*Does not clarify question or check response.* |
| **2.0 ~ 2.5** | *Conscious of turn-taking*  
*Responds to others but often with unnatural gaps or pauses*  
*May show some basic turn-taking but does not relate ideas well, or give much explanation* | *Shows limited organization of information*  
*Communicates basic information about the visual prompt, but omits numerous significant details*  
*Description contains frequent inaccuracies* | *Shows general comprehension of question.*  
*Response addresses question but is not very effective.*  
*May not clarify or check question.* |
| **3.0 ~ 3.5** | *Initiates turn taking*  
*Maintains interaction by responding to others without unnatural gaps or pauses*  
*Shows meaningful agreement or disagreement to others’ opinions (assent / dissent, etc)* | *Shows some attempts at organization of information*  
*Can communicate a range of information about the visual prompt, but makes some omissions*  
*Description contains some inaccuracies* | *Shows clear comprehension of question.*  
*Response answers question fairly effectively.*  
*Clarifies or checks question.* |
<table>
<thead>
<tr>
<th>Score</th>
<th>Features</th>
</tr>
</thead>
</table>
| 4.0 - 4.5 | - Appears confident  
- Responds appropriately to others  
- May direct conversation  
- Shows ability to negotiate meaning quickly and naturally  
- May begin to use paraphrase or clarification  
- Shows good organization of information  
- Communicates nearly complete information about the visual prompt, with minor omissions  
- Description has minor inaccuracies  |
| 5.0 | - Very confident and natural  
- May ask others to expand on views  
- Negotiates, holds and relinquishes turns appropriately  
- Explains how own and others’ ideas are related, interacts smoothly  
- Shows excellent organization of information  
- Communicates complete information about the visual prompt  
- Description contains few or no inaccuracies  |
|  | - Content mostly well-developed; some minor ideas may be irrelevant or repetitious  
- Information mostly organized; occasionally omits transitions or supporting evidence  
- Delivery is mostly acceptable, with minor problems  |
|  | - Shows clear comprehension of question.  
- Response answers question effectively.  
- Clarifies and checks question.  |

- Content relevant to the topic and well-developed  
- Organization is well demonstrated throughout  
- Delivery is acceptable as an academic presentation  
- Shows clear comprehension of question.  
- Response answers question very effectively.  
- Clarifies and checks question.  |
Appendix C Example Picture and Graph description task

Hours Worked at Job Each Week

<table>
<thead>
<tr>
<th>Job</th>
<th>Number of Hours</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cook</td>
<td>36</td>
</tr>
<tr>
<td>Manager</td>
<td>50</td>
</tr>
<tr>
<td>Banker</td>
<td>25</td>
</tr>
<tr>
<td>Doctor</td>
<td>80</td>
</tr>
<tr>
<td>Nurse</td>
<td>65</td>
</tr>
</tbody>
</table>