Designing Speaking Tests for Freshman English Classes:

Looking to the Commercial Sector for Inspiration

Adrian Wagner, Asia University

Abstract

Teaching *Freshman English* classes is a major part of the teaching load for visiting faculty members at Asia University. Beyond a designated percentage for attendance and participation, teachers are free to design assessment and allocate grades as they see fit. Teachers seeking to use speaking tests to generate a portion of students' grades face many challenges in designing and administering speaking tests. This essay seeks to inform and assist Freshman English teachers in creating speaking tests through assessing the formats of three well-known commercial speaking tests, namely the TOEIC Speaking Test, the IELTS speaking test and the Cambridge suite of exams. Particularly it investigates elements such as the length of the tests, styles of questions and grading schemes. Hopefully, Freshman English teachers will find this essay useful when they are preparing and administering speaking tests in their classes.

Introduction

Communication-based English classes, especially as required classes for first year students, are prevalent in Japanese universities. And as communicative language teaching as a concept and a practice makes more and more inroads into the consciousness of foreign language education, more and more teachers are emphasising communicative ability and incorporating speaking activities such as pair and group work to increase student speaking time in class. With in-class activities and educational goals focusing on communication skills, many teachers feel that at least a portion of the overall grade of the student should be based upon their speaking performance. Disregarding assessment tasks such as presentations and ongoing assessment, creating and administering a speaking test seems the most simple and most straightforward method of being able to justify the giving of a grade for speaking.

Implementing speaking tests often proves challenging for a number of reasons. Firstly, practical issues such as logistics and time need to be considered. At Asia

University, most Freshman English classes contain between 20 and 25 students with a class time of 45 minutes. Were a teacher to attempt to give an individual speaking test to every student in the class during a single class period, it would allow for less than two minutes per student.

Aside from logistical considerations, we must consider theoretical issues, such as scoring criteria and the inherent subjectivity of giving a number, percentage, grade or other type of quality judgment to a student's speech.

Facilitating and administrating speaking exams is time and labour intensive. Unlike the multiple choice question and answer styles of the listening and reading test that can be scored easily, speaking and writing tests require the tester to spend time both administering and marking the tests. While computer administered testing has been utilised for some time, even for speaking and writing tests, the assessment of tests, whether we call it rating, scoring, marking or grading, still requires the direct time and attention of a skilled and trained human or humans.

Outside of universities and institutions, English assessment is big business. With considerable resources to use and a vested interest in providing tests and results that are seen as valid while being as human resource efficient as possible to remain commercially competitive, the world of commercially available speaking tests potentially have a lot to offer classroom teachers struggling with the question of how to administer and assess speaking test efficiently and consistently. This essay will investigate three well-known, internationally-recognised speaking tests, investigate their formats and assess their suitability for adaptation by Freshman English teachers at Asia University.

The TOEIC Speaking Test

The TOEIC Speaking Test is a computer based test in which the examinee's voice is recorded. The test is taken individually, takes about 20 minutes and is marked on a score scale of 0-200. Table 1 is a summarised version of the test contents. (ETS, 2008, p6.)

Table 1: TOEIC Speaking Test Question Types

Question Number	**Question Type**
1-2	Read a text aloud
3	Describe a picture
4-6	Respond to questions
7-9	Respond to questions using information provided
10	Propose a solution
11	Express an opinion

Preparation time is given for questions one, two, three, ten and eleven. To give an indication of the range of difficulty level of questions, the following are examples taken from a sample test provided by ETS. Questions in the third section of the test, (questions 4, 5 and 6) begin relatively simply but demand more detailed answers as they progress.

Question 4: How often do you watch television?
Question 5: What kinds of programs do you usually watch?
Question 6: Describe your favorite television program. (ETS, 2008, p. 7)

Questions 7-9 are based around a text. The examinee is given time to read the information and then asked practical questions about it.

Question 10 is perhaps the most challenging of the tasks as it requires the examinee to listen to a spoken dialogue explaining a problem before proposing a solution.   Question 11 is also very challenging as it requires someone to express and support and opinion.

As stated above, the TOEIC Speaking Test is marked on a holistic scale. "In holistic scoring, one or more readers assigns a single grade or rating to a text, based on an overall, total impression" (Terry, 1986 p. 525). In the case of a speaking test, the aforementioned, "text" could be interpreted as representing the whole answer to a question.

Discussing the use of holistic scales in assessing writing, Liskin-Gasparro and Woodford (as cited in Terry, 1986, p. 127) propose two basic concepts, (1) “the whole of the essay is greater than the sum of its parts, and (2) teachers who are experienced with writing can recognize good writing when they see it, even if they cannot come to an agreement on how to describe it!” These concepts have been directly applied to the grading of speaking, although obviously the “parts” and “whole” are samples of speech rather than paragraphs or essays.

Many researchers have highlighted the limitations of holistic rating systems. “The main problems in the use of holistic rating concern validity: what a holistic score actually represents and whether certain aspects outweigh others as assessors to form an overall judgment of test-taker performance” (Iwashita and Grove, 2003 p.26). The validity that these researchers appear to question is the focus of this kind of system. With such a broad and overarching system, it is difficult to identify what exactly is being assessed and how. Another concern is the reliability of holistic marking scales. Would a different examiner give the same or similar score to a test-taker? Would the same instructor assign the same score if they heard the same response from a different person or from the same person at a different time of day? To some degree, we can counteract the subjectivity of holistic rating systems by being clear with ourselves and for the sake of fairness, with the students as well, about which, if any skills or criteria will be prioritised in the assigning of a score. Also, it must be stated that not only holistic scoring scales, but all scales used to rate language ability will suffer from doubts about reliability and validity.

Another complication in using holistic scoring scales comes from the infinite number of responses that a single question could generate. “Certain components in free creative responses cannot be quantified as discreet-point items since there can be no clear-cut anticipated response” (Terry, 1986, p.525). Teachers using a holistic scale to assess speaking need to be clear and realistic with themselves regarding what they would consider to be an excellent, good, average or bad answer. A teacher using a holistic scale could base these expectations on such factors as the level of the class, the content of the class and the materials used.

Despite the inherent subjectivity of holistic schemes, they are still widely used, so there must be some advantages. Becker (2010) summarises three main advantages as:

that it emphasises what is done well, represents a direct reaction to and impression of the produced language and is faster and simpler. The final consideration would probably be the most attractive to freshman English teachers, as they finish one exam, check their watch and look at the long list of students' names waiting their turn to be tested.

For the TOEIC test, questions are rated individually. "The first four task types (Questions 1–9) are rated on a scale of 0 to 3 and the last two task types (Questions 10–11) are rated on a scale of 0 to 5" (ETS, 2010, p.9). Questions 10 and 11 are seen as being more difficult and therefore have a wider scoring range.

Examinees taking the speaking test are given a proficiency rating in of one to eight which corresponds to a total score out of 200. This score is a mathematical function of the rating system of individual questions above. For example, the descriptor for the highest level of achievement is below:

> Level 8 Scale Score 190–200
>
> Typically, test takers at Level 8 can create connected and sustained discourse appropriate to the typical workplace. When they express opinions or respond to complicated requests, their speech is highly intelligible. Their use of basic and complex grammar is good and their use of vocabulary is accurate and precise. Test takers at Level 8 can also use spoken language to answer questions and give basic information. Their pronunciation, intonation, and stress are at all times highly intelligible. (ETS, 2010, p.11)

From the perspective of a classroom teacher at a university, it would seem that the eight level scoring system, and particularly the descriptors employed by the TOEIC system, would not offer much practical use for in-class speaking tests. In the Freshman English program at Asia University, presumably, as all students have been streamed and assigned roughly corresponding levels by the in-house placement test, the students in a single class would (theoretically) at the beginning of the course have abilities in a similar tier. Basically, the rating system employed by the TOEIC course, utilised as it is, would be too broad to measure the achievement of Freshman English students over one semester or even one year. Finally, allocating twenty minutes per student would be difficult or impossible. In a class of twenty students, this would require almost seven hours of class time.

However, some aspects of the TOEIC Speaking Test and rating design could be useful to consider when creating speaking assessment for Freshmen English tests. The range of question difficulty, the use of the holistic scale and assigning different levels of possible full marks per question and tallying these to give an overall total, offer a realistic model that could be applied in the freshman English classes. Furthermore, the range of question styles, incorporating direct interview styles, picture stimuli and written stimuli, gives teachers options to adapt tests to suit the material used in class.

The IELTS Speaking Test

The IELTS Test is another major English test which is the standard for university entrance for speakers of English as a second language in places such as the United Kingdom and Australia. The IELTS test incorporates reading, writing, listening and speaking skills and is different to the TOEIC test in that tests for receptive and productive skills cannot be taken separately. Unlike the TOEIC Speaking Test which is now taken via a computer, the IELTS test is taken in person with the examinee responding to questions face to face with an examiner.

Similar to the TOEIC test, IELTS speaking test results are given to examinees as a proficiency rating of one (lowest) to nine (highest). There is a 0 score given if the examinee does not attend.

In terms of time, the test is shorter than the TOEIC test, taking between 11 and 14 minutes per examinee. It contains three sections. The first section is an interview in which the examinee is asked basic questions about familiar issues. It is designed to “help you relax and talk naturally” ("Take Ielts with British Council," n.d.).

In the second part of the test, the examinee is given a task card and, after one minute of preparation time, is expected to speak about the topic on the card for between one and two minutes without interruption. A sample task card from IELTS (n.d.) is attached as Appendix A.

After finishing speaking, the examinee is asked one or two questions about the topic. Finally, in the third part of the exam, the examinee is asked some questions related to the topic from the previous section of the test, which give the test taker an opportunity to expand upon the topic and speak about more abstract ideas.

A key difference between the TOEIC and the IELTS is that the IELTS utilises an analytic scale for rating the test with separate scores given for different elements of language, which are averaged to an overall score. Analytic (also referred to as analytical) scoring schemes assess individual components of a whole. As opposed to the holistic scale, which gives a single score or rating based on an overall impression, the analytic scale is explicit in providing a different score for various aspects of language use. For example, an analytic rubric that many teachers may be familiar with is the one provided in the *Interchange* (Richards, 2012) series of textbooks. The *Interchange* rubric for oral tests has the separate elements of comprehension, fluency, grammar, vocabulary and pronunciation. It is provided as Appendix B.

In the IELTS Speaking Test, the four individually rated categories are fluency and coherence, lexical resource, grammatical range and accuracy, and pronunciation.

There are advantages attributed to analytic schemes. At the most basic level, the outwardly more mathematical system should provide greater accuracy. "The evaluation criteria are more focused, grading tends to be more reliable" (Terry, 1986. p 525). The person grading the test has explicit criteria and the ability to give and individual rating for each component. Ideally, this should make scoring more consistent.

From the perspective of language learners, this type of scheme provides more useful feedback than the scores generated from holistic scales. Tuan (2012) writes, "it provides more useful diagnostic information about students' speaking abilities. That is, it tells learners where their weaknesses are and where their strengths are" (p 674). While people taking the IELTS test are not shown the rubric used when receiving their results as a proficiency rating discussed above, a classroom teacher would have the option of returning these rubrics, or relevant information gathered from the rubrics, to the students. This data could be analysed and inform classroom practice by identifying trends amongst a group of students.

While adding some comfort and convenience, and some sense of order to fall back on, dividing the components of language also brings its own complications. It brings into question how we separate speech and judge its elements distinctly. "…although analytic scoring may improve reliability among measurements, the scoring of one individual trait can influence how another trait is scored" (Becker, 2010, p. 113). Presumably, this quote refers to not only the reliability of the test but also the validity. Are we really scoring what we think we are scoring? These types of schemes certainly do not negate the possibility of conflation or other types of biases.

It is virtually impossible to remove the element of subjectivity when assessing speaking, no matter how specific the individual elements on the rubrics may be. In a study assessing the successes and shortcomings of the four scale analytic system currently used in the IELTS speaking test, the researcher noted *fluency and coherency* as the item which contained the most ambiguity and provided the most difficulty to the examiners, as it was covering "a larger number of relatively discrete aspects of performance than the other scales– hesitation, topic development, length of turn, and use of discourse markers" (Brown, 2000, p 8). Use of analytic systems cannot remove the difficulty of making personal and inherently individualistic judgments. Trained IELTS examiners remarked on the difficulty of reacting to hesitation in an examinees speech:

They also frequently attempted to infer the cause of hesitation, at times attributing it to linguistic limitations… to their personality (shyness), to their cultural background, or to a lack of interest in the topic… Often examiners were unsure whether language or content was the cause of disfluency but, because it was relevant to the ratings decision they struggled to decide. (Brown, 2000 pp. 8-9) Converse to its intended purpose, the analytic scale in this case is seemingly complicating the rating process.

Analytic schemes are seen as highlighting the negative, as opposed to the holistic schemes which focus on the strengths of the language being examined. "The analytic scoring often has the tendency to reduce and oversimplify the components of speaking, and to emphasize the flaws rather than the strengths of speaking" (Tuan, 2012 p. 674). It is possible that when over-analysing individual aspects of a students' speech, we will lose sight of the bigger picture and the basic purposes of language.

Another factor to consider when deciding on a scoring scale is that generally, analytic scales are seen as being more time-consuming. They can be distracting to the examiner while giving the test and take more time and effort to tally the individual scale scores after the exam has been completed.

Perhaps, the aspect of the IELTS test design that could be most easily incorporated into Freshman English speaking tests is the second section of the test, in which a student is instructed to speak at length about a topic without interruption. Giving the student a task card, some preparation time and a time target is a good way to ensure that a larger amount of uninterrupted speech can be elicited and examined.   The complexity of the topics on the cards, follow up questions and target time can be adjusted to suit the level of the class.

While it does have its flaws and imperfections, an analytic scoring scheme is a useful option if the teacher hopes to give students detailed feedback about the speaking test result. Whether these would be motivating or demotivating to the students would depend on the individual and is another issue entirely.

The Cambridge Suite

The Cambridge exams differ from the TOEIC and the IELTS in a number of ways. There are various levels of exams available. Learners choose an exam to suit their level and will be given a pass or fail result. The main general English exams, beginning with the easiest, are *Key English Test* (KET), *Preliminary English Test* (PET), *First Certificate* (FCE), *Advanced* (CAE), and *Proficiency (*CPE). All of the exams have a different format. Tables briefly outlining the contents of the KET exam and the PET exam (the levels probably best suited to freshman English students at Asia University) are provided as Appendices C and D.

The greatest difference between the Cambridge tests and most other speaking tests are that the Cambridge test is not taken individually but in groups of two or three examinees, who take the exam together. It is tempting to dismiss this decision as being one born out of convenience and commercial considerations. However, the literature available provides both advantages and disadvantages to this kind of testing and does provide some legitimate reasons for testing students in pairs or small groups rather than individually.

Referring to the research of Saville and Hargreaves (1999) Norton writes that, "candidates are more relaxed; they have the possibility of more varied patterns of interaction during the tests; and this format can lead to positive washback in the classroom by encouraging learners to interact together in preparation for the test" (2005, p. 288). The factor of encouraging cooperation between learners in the classroom, prior to the test, could be a particularly strong point in favour of adopting such a test structure. In the Freshman English classes, teachers are always seeking to motivate their students to communicate in English with each other, and a direct relationship of similarity between in-class activities and the test itself could help to increase student effort in communicative tasks.

Writing in response to criticism by detractors of the paired format of the Cambridge suite of exams, Taylor (2003) also sites classroom interaction as a reason for moving away from the individual format. The decision to change was, "based on pedagogical considerations (i.e. a more communicative approach to language teaching with the use of pair/group work in the classroom)" (p. 15). The assertion that classroom activities and assessment activities should reflect each other to some degree is a compelling one. If our classes are based on pair and group work, a one-to-one interview between student and teacher is a different task to what the students have become accustomed to. Is it fair to assess them for something they have never really done in class?

Other support for this type of test format is based on theories of second language discourse. It is posited that in one-to-one interview tests, through the asymmetrical relationship between examiner and examinee, the range of language functions is limited and is largely limited to informational functions. In adding another examinee to the exam, this lack of balance is somewhat mitigated, allowing for other types of language functions to be used and assessed. “Overall, the range of functions elicited by the paired speaking test format proved to be much greater that the one-to-one format (26 functions out or 30 for the paired format and 14 out of 30 for the one to one” (Taylor, 2003, p. 17). She argues that the paired format encourages greater use of interactional functions, among others.

Support for this theory also comes from analysis of the IELTS test. In comparing aspects of language assessed in the IELTS speaking test, compared to the linguistic demands of actually attending university classes, the authors observed that, “In classroom interaction, students are also involved in the production of a range of interactional and interaction management functions, whereas they are not required, but may occur, in the IELTS interview” (Ducasse and Brown, 2009 p. 19). It should be noted that depending on the level of the Freshman English class and the textbook used, these interactional functions may be beyond the scope of the class and abilities of the students, making this somewhat irrelevant.

Another key difference in the format of the Cambridge speaking tests is that there are two examiners present, an assessor and an interlocutor. The assessor rates the examinees using an analytic scale and the interlocutor rates using a holistic scale. This aspect of the exam could not be easily replicated in Freshman English classes and teachers may have difficulty rating two or more students at a single time while conducting the exam by themselves.

Criticism and complications of the test appear to be centred on how different pairings affect the linguistic performance of the individual. Factors such as gender, how well the candidates know each other, as well as differences in personality and confidence need to be considered. Norton (2005) observes that, “being paired with a candidate who has higher linguistic ability may be beneficial for lower level candidates who are able to incorporate some of their partner’s expressions into their own speech” (p. 287). Having to decide whether a student has learned and understands the language they are using or is simply mimicking certainly complicates the decisions involved in the grading process.

For Freshman English teachers at Asia University, the arguments in favour of paired or small group speaking tests would seem to be based on considerations of time and establishing a similarity between classroom and assessment tasks. The complications of social interaction and the demands of assessing more than one student simultaneously would seem to be reasons against emulating aspects of the Cambridge paired speaking test format.

Conclusion

There is no *best way* to administer and rate speaking tests. All structures, question types and rating schemes have both their advantages and limitations. The analysis of three widely taken and accepted speaking tests above should provide an overview of some popular methods for grading and administrating speaking tests, and some aspects that could be incorporated into assessment design in Freshman English classes. When teacher at Asia University compare these different tests, they may find themselves pulled in different directions based on their own academic, philosophic and pragmatic inclinations. As educators, we should always be refining our practices, questioning why we are doing things a certain way and seeking alternatives and improvements. In designing oral assessment we have a lot of options, and hopefully the information provided in this paper will help us know those options a little better.

References

Becker, A. (2010). Examining rubrics used to measure writing performance in US intensive English programs. *The CATESOL Journal*, *22*(1), 113-130. Retrieved from http://www.catesol.org/Becker%20113-130.pdf

Brown, A. (2000). An investigation of the rating process in the IELTS oral interview. *IELTS research reports*, *3*, 49-84. Retrieved from https://www.ielts.org/pdf/Vol3Report3.pdf

Cambridge English. (2014). General English and for schools. Retrieved from http://www.cambridgeenglish.org/exams/general-english-and-for-schools/

Ducasse, A. M., & Brown, A. (2011). *The role of interactive communication in IELTS Speaking and its relationship to candidates' preparedness for study or training contexts* (Vol. 12, pp. 1-26). IELTS Research Report. Retrieved from http://www.ielts.org/pdf/Vol12_Report3.pdf

IELTS. (n.d.) *Speaking sample task part 2.* Retrieved from https://www.ielts.org/pdf/115047_Speaking_sample_task_-_Part_2.pdf

Iwashita, N., and Grove, E. A comparison of analytic and holistic scales in the context of a specific purpose speaking test. *Prospect*, *18*(3,) 25-35. Retrieved from http://www .ameprc.mq.edu.au/docs/prospect_journal/volume_18_no_3/18_3_2 _Iwashita.pdf

ETS. (2008). *TOEIC Speaking and writing practice tests.* Retrieved from https://www.ets.org/Media/Tests/TOEIC/pdf/TOEIC_sw_sample_tests.pdf

ETS. (2010). *TOEIC User guide speaking and writing* . Retrieved from http://www.ets.org/s/toeic/pdf/toeic_sw_score_user_guide.pdf

Norton, J. (2005). The paired format in the Cambridge Speaking Tests. *ELT Journal*, *59*(4), 287-297. Retrieved from http://203.72.145.166/elt/files/59-4-2.pdf

Richards, J. C. (2012). *Interchange Intro Teacher's Edition with Assessment Audio CD/CD-ROM*. Cambridge University Press.

Taylor, L. (2003). The paired speaking test format: Recent studies. *Perspective*, *55*(1), 70-76. Retrieved from http://www.standardsicherung.schulministerium.nrw.de/cms/upload/Sprachenwerkstatt/ResearchNotes6PairedSpeakingTests.pdf

Terry, R. M. (1986). Testing the productive skills: A creative focus for hybrid achievement tests. *Foreign Language Annals*, *19*(6), 521-528.

Tuan, L. T. (2012). Teaching and Assessing Speaking Performance through Analytic Scoring Approach. *Theory and Practice in Language Studies*, *2*(4), 673-679.

Appendix A

IELTS Speaking Test Sample

Part 2 – Individual long turn

Candidate Task Card **Describe something you own which is very important to you.**
You should say:
where you got it from
how long you have had it
what you use it for
and explain why it is important to you.

You will have to talk about the topic for 1 to 2 minutes.
You have one minute to think about what you're going to say.
You can make some notes to help you if you wish.

Rounding off questions

- **Is it valuable in terms of money?**
- **Would it be easy to replace?**

(IELTS, n.d.)

Appendix B

Interchange Oral Quiz Scoring Sheet

Oral quiz scoring sheet Name: ___________________________

Date: ________________________

Score:________________________

		Poor	**Fair**	**Good**	**Very good**	**Excellent**
Comprehension	**0**	**1**	**2**	**3**	**4**	**5**
Fluency	**0**	**1**	**2**	**3**	**4**	**5**
Grammar	**0**	**1**	**2**	**3**	**4**	**5**
Vocabulary	**0**	**1**	**2**	**3**	**4**	**5**
Pronunciation	**0**	**1**	**2**	**3**	**4**	**5**

General comments

Suggestions for improvement

____

____

(Richards, 2012)

Appendix C

Cambridge Key Speaking Test Summary

Part 1 (Interview)

What's in Part 1?	Conversation with the examiner. The examiner asks you some questions about yourself and you answer.
What do I have to practise?	Giving information about yourself.
How long do we have to speak?	5–6 minutes

Part 2 (Collaborative task)

What's in Part 2?	The examiner gives you some information or a card with some ideas for questions. You have to talk with the other candidate and ask or answer questions.
What do I have to practise?	Asking and answering simple questions about daily life.
How long do we have to speak?	3–4 minutes

(Cambridge English, 2014)

Appendix D

Cambridge PET Speaking Test Summary

Part 1 (Interview)

What's in Part 1?	Conversation with the examiner. The examiner asks questions and you give information about yourself, talk about past experiences, present job, studies, where you live, etc., and future plans.
What do I have to practise?	Giving information about yourself.
How long do we have to speak?	2–3 minutes

Part 2 (Discussion)

What's in Part 2?	The examiner gives you some pictures and describes a situation to you. You have to talk to the other candidate and decide what would be best in the situation.
What do I have to practise?	Making and responding to suggestions, discussing alternatives, making recommendations, negotiating agreement.
How long do we have to speak?	2–3 minutes

Part 3 (Extended turn)

What's in Part 3?	The examiner gives you a colour photograph and you have to talk about it.
What do I have to practise?	Describing photographs.
How long do we have to speak?	3 minutes in total; 1 minute to talk about the photograph.

Part 4 (General conversation)

What's in Part 4?	Further discussion with the other candidate about the same topic as the task in Part 3.
What do I have to practise?	Talking about your opinions, likes/dislikes, experiences, habits, etc.
How long do we have to speak?	3 minutes

(Cambridge English, 2014)