Tips for Testing Speaking
by [Carla Schnitzler Hall](#) and [Amelia Kreitzer Hope](#)

Assessing speaking skills is challenging for a number of reasons. It's time consuming, and, unlike writing, speech is ephemeral—it's there and then it's gone. Even if a speech sample is recorded, it is often difficult to identify the individual strengths and weaknesses that characterize a test-taker's overall speaking abilities. There are, however, some general guidelines that can make the job of assessing oral skills easier. In this article we will discuss two key areas of testing speaking: 1) eliciting a rich, representative speech sample, and 2) evaluating the sample reliably.

Eliciting a Good Sample

Test Pairs or Groups of Students

One way to make testing speaking easier is to use a test format that evaluates more than one student at a time. In addition to saving time, testing pairs or groups of students at the same time often elicits conversational language and interactional skills (e.g., disagreeing, interjecting, and reprimanding) that are difficult to tap in a traditional, one-on-one interview format test with an examiner only. Another advantage of this type of testing is that the interaction takes place between the test takers, leaving the examiner free to focus on the assessment only. Paired or group testing can also lead to positive classroom washback if the task the test takers engage in is based on classroom work.

In paired or group assessments, it is important to make sure there is a clear task for the test takers to complete. Such tasks could include:

- problem-solving tasks
- information gap exercises
- consensus exercises
- role-plays
- debates
- discussions
- oral presentations

If you decide to assess pairs or groups of students at the same time, be aware that these types of speaking tasks often require some work to develop, and they should be field-tested to ensure that the task actually provides opportunities for test takers to produce extended responses. Also consider that in paired or group tests, shy or quiet test takers may participate less, and, as an examiner, you may need to intervene in the assessment to make sure that each test taker contributes enough language for you to rate.

Use Technology

Another way to simplify obtaining a speech sample is to use readily available recording technologies. If you do not have access to a computer lab or digital recorders, students can record their voices on their own computers or use their smart phones and send the files to you. Whether

this is done during or outside of class, it greatly reduces the amount of class time it takes to gather the samples.

Conduct Interviews

If you decide to conduct one-on-one assessments, there are a number of formats you can use. The most traditional type of examiner–test taker interaction is the interview. An interview can be scripted, where all or most of what the interviewer says and asks the test taker to do is prepared and written down, or it can be unscripted. Scripted and semiscripted interviews are more consistent and reliable, and they allow the examiner to focus on evaluating the speech sample more than on the interview process. They can, however, involve a lot of work to prepare, and they may result in a less realistic conversation.

Unscripted interviews take less time to prepare and allow the examiner, and potentially the test taker, to move in unexpected directions, possibly producing a more authentic interaction and opportunities to test a wider range of test taker linguistic resources. The trade-off is that the interviewer must concentrate on both the interview process and on evaluating the test taker's performance at the same time. Additionally, the quality of any interview, especially in unscripted interviews, is heavily dependent on the skill of the interviewer. Unscripted interviews may, therefore, require more training and monitoring of interviewers.

Use a Prompt

Interviews can also be based on a prompt where test takers complete, look at, read, or listen to a stimulus and then respond to a prompt based on that stimulus. One possible prompt might be an information form that the test taker completes before the interview starts. The form could ask questions about the test taker's place of birth, educational background, work experience, hobbies and interests, goals for the future, or any other relevant personal information. What is nice about using a prompt like this is that it gives the test taker some control over the topics discussed in the test and provides the interviewer with topics that are germane to the test taker. As with any test prompt, the interviewer can build on the information obtained from the form and the initial interaction and move on to related but different topics of discussion.

Pictures or pictures series are another type of prompt that can be used to start a discussion. One benefit of using pictures is that they do not rely on the test taker's ability to listen or read, so they are suitable for all levels of proficiency. When choosing pictures, make sure that they are clear enough to decode, they are not too culturally specific, they tell a story or imply a back story, they are visually dense (i.e., they include sufficient visual information so that there is something to talk about), and they are a good springboard for further discussion. (See Appendix A for a sample picture series prompt.)

Create Good Questions

Regardless of the prompt you use, ensure that you provide ample opportunity for test takers to display their full range of linguistic resources. To do this, create questions that challenge or push them to the limits of their abilities. One way to do this is to think about the grammar the question

is likely to elicit in the response. For example, if you are using a picture prompt, you may ask the test taker to describe the picture to elicit simple and progressive present tenses. You could follow up by asking the test taker to speculate on what happened before the picture was taken to elicit past tenses, and then ask the test taker to predict what will happen next to prompt the use of modals or future tenses. You could also ask hypothetical questions in any tense to create opportunities for the test taker to use conditionals.

Also consider the language functions that your questions might elicit. To tap narration, you could ask test takers to recount a story. You could also create questions that require test takers to compare, explain, hypothesize, or defend a position. Role-plays can often provide opportunities for test takers to use functions that are difficult to elicit in other test tasks. A well-constructed role-play could place the test taker in the position of having to give advice, make a complaint, apologize, reprimand, or express a need.

Keep in mind that your goal is to generate a rich, representative sample of the test taker's abilities, so you may need to tailor your questioning to the individual abilities of test takers. You can increase or decrease the difficulty of questions by adjusting their length, their grammatical complexity, and their vocabulary level. Questions that require test takers to talk about abstract concepts often require test takers to use more complex language than do questions about personal or concrete matters. Also keep in mind that open-ended questions will generally elicit longer responses than yes/no questions. (See Appendix A for sample questions accompanying a picture series prompt.)

Rating the Sample

Use a Rating Rubric, Checklist, or Scale

Just as important as obtaining a good sample is rating that sample reliably. Rating rubrics, checklists, or scales are tools that can help you do this. Each of these tools lists criteria for a piece of work or performance and also articulates gradations of quality for each criterion. They usually include criteria related to accuracy and range in grammar and vocabulary, pronunciation, fluency, listening skills, and overall comprehension. Here are links to some good sample rubrics:

- Independent Speaking Rubrics (ETS)
- PALS: Performance Assessment for Language Students (Fairfax County Public Schools)
- IELTS Speaking Band Descriptors (British Council, IELTS Australia, University of Cambridge ESOL Examinations)

Good rating rubrics, checklists, and scales are tied directly to the task that is being assessed. So, for example, if the task requires test takers to interact conversationally, the rating tool should measure success at interaction. If, on the other hand, the task is to make a presentation, the rating tool may measure the test taker's use of visuals and body language. Good rating tools also use terms that are shared and understood by all raters. If the tool uses the term "register," for instance, all raters should know what that term means, and they should all evaluate it in the same way. Finally, good rating tools are manageable. Usually four or five criteria (e.g., accuracy,

pronunciation, fluency, etc.) and four or five levels (e.g., limited, low-intermediate, intermediate, etc.) are comprehensive enough while at the same time remaining user friendly.

Using these types of rating tools has advantages for both teachers and students. They can help teachers clarify and explain expectations, thus enhancing accountability. In well-designed rating tools, specific performance features are clearly outlined, so they are easy to use and make grading quicker. They are also flexible in that they can be modified to fit the level of the students you are assessing and the task you are using. Rating tools also provide students with clear expectations and criterion-referenced feedback. If appropriate, teachers can involve students in the rubric design process, adding a sense of goal setting and personal responsibility to the evaluation process.

Ideally, rating tool descriptors should be derived from many real speech samples. When designing rating tools, try to avoid using negative or comparative (norm-referenced) language. For example, instead of using phrases such as "level B test takers are more fluent than level A test takers," develop descriptors that actually describe test takers' behaviours at each level, such as "level A test taker's speech is characterized by frequent hesitations and false starts." Be sure that descriptors include all the characteristics of the performance you want to rate and that they describe behaviours that you can actually observe. Descriptors should also clearly discriminate between levels. Finally, if the rating tool is used by different examiners, periodic monitoring should take place to ensure that all raters are applying the rating tool in the same way.

Rate Speech Samples in Real Time

Another way to increase efficiency in testing speaking is to score test-taker performances in real time (whether they are live or recorded) instead of stopping and starting or relistening to recordings of the speech samples. Again, good rubrics, checklists, and scales lend themselves to quick scoring because raters simply need to circle or highlight the descriptors or scale number that applies to the test taker's performance.

Although assessing speaking skills may not be as easy as evaluating other areas of second language proficiency, there are ways to make the job less onerous. Developing efficient methods of eliciting speech samples and creating easy-to-use rating tools ensure that you obtain a representative of test takers' abilities and that you score that sample reliably.

———————————————————

Carla Hall is a language teacher at the Official Languages and Bilingualism Institute at the University of Ottawa.

Amelia Kreitzer Hope is head of Language Testing Services at the Official Languages and Bilingualism Institute at the University of Ottawa.